MASTER EN SCIENCES MATHÉMATIQUES

Application de la méthode divisive de classification symbolique à des données modales

Douny, Isabelle

Award date:
1999

Facultés Universitaires Notre-Dame de la Paix
Namur
Faculté des Sciences - Département de Mathématique

Application de la méthode divisive de classification symbolique à des données modales

Mémoire présenté pour l'obtention du grade
de Licencié en Sciences
mathématiques
par

Isabelle DOUNY

Promoteur : J.-P. RASSON

Année académique : 1998-1999

Je tiens à remercier, tout particulièrement Monsieur Jean-Paul Rasson pour son aide, sa disponibilité et sa gentillesse tout au long de ce travail.

Je tiens à exprimer ma reconnaissance à Madame Sandrine Baudart-Lissoir, Monsieur Benoit Colson, Mademoiselle Alexandra Dessy et Monsieur Robert De Mol pour leur soutien, leurs encouragements ainsi que leurs conseils. Je tiens également à remercier Messieurs Marc Dubuisson et Jean-Paul Duprez, membres du Service des études et de la Statistiques du Ministère de la Région Wallonne pour leur aide lors de l'interprétation des résultats et pour la réalisation des cartes de Wallonie.

Mes pensées se tournent également vers ma famille et mon entourage pour leurs encouragements et leur patience durant ces quatre années d'études.

Enfin, je dédie ce travail de fin d'études à mon grand-père.

A tous, merci

Résumé

Ce mémoire a pour but de tester une méthode divisive de classification symbolique réalisée dans le cadre du projet européen SODAS.

L'objectif de cette méthode est de décomposer un ensemble de données de n objets décrits par un ensemble de p caractéristiques, en un nombre de classes d'individus "semblables" en utilisant les objets symboliques.
Nous appliquerons cette méthode à des fichiers dont les variables modales caractérisent les communes wallonnes.

Abstract

The subject of this work is to test a divisive method of symbolic clustering realized as part of the SODAS european project.

The aim of this method is to decompose a given set of n objects described by a set of p features in a number of clusters of similar objects.
We will apply this method to different files. Their modal variables describe the "wallonnes" municipalities.

Table des matières

I La partie théorique 8

1 Le projet SODAS (Système officiel de l'Analyse des Données symboliques) 9

1.1 Les objectifs 9
1.2 Le logiciel-prototype SODAS 10
1.3 Les méthodes d'Analyse des Données symboliques 10
1.4 Le plan de travail du projet 11
1.4.1 La construction, la gestion et la manipulation des objets symboliques 12
1.4.2 Les méthodes de l'Analyse des Données symboliques . . 12
1.4.3 Lien avec l'analyse des données classiques 14
1.4.4 Le rapport scientifique 15
1.4.5 L'utilisation 15

2 Rappels de méthodes classiques de statistiques 16

2.1 Introduction 16
2.2 Les variables classiques 16
2.2.1 Les variables quantitatives 17
2.2.2 Les variables qualitatives 18
2.3 La matrice des données 21
2.4 Les données manquantes 22
2.5 Exemple 23

3 Les données symboliques 24

3.1 Introduction 24
3.2 Préliminaires 24
3.3 Variables définies dans un ensemble de valeurs 25
3.3.1 Définitions 25

3.4 Variable modale 28
3.4.1 Préliminaires 28
3.4.2 Définitions 29
3.5 Le tableau des données symboliques 31

4 Les objets symboliques **36**
4.1 Introduction 36
4.2 Relations et descriptions 38
4.2.1 Introduction 38
4.2.2 Rappelons la terminologie utilisée pour les relations . . 38
4.3 Objets assertions 40
4.3.1 Définitions 40
4.4 Objets symboliques individuels 43
4.4.1 Définitions 43
4.5 Objets symboliques booléens 45
4.5.1 Introduction 45
4.5.2 Définitions 46
4.6 Définitions générales d'objets symboliques - Objets Modaux . 47
4.6.1 Préliminaires 47
4.6.2 Définitions générales 48

5 La similarité et la dissimilarité **49**
5.1 Les mesures classiques de similarités et de dissimilarités 49
5.1.1 Introduction 49
5.1.2 Définitions 49
5.1.3 Mesures de ressemblance entre objets 50
5.2 Les fonctions-distance et propriétés spéciales 52
5.2.1 La matrice de dissimilarités 52
5.2.2 Mesures de distance à partir d'une matrice de données classiques 54
5.2.3 Mesures de similarité pour les matrices de données-catégories 57

6 Les méthodes de classification pour des objets symboliques **62**
6.1 Le problème de classification et des méthodes de classification pour des données classiques 62
6.1.1 But de la méthode 62
6.2 Rappels de concepts de base (pour les données classiques) . . . 63

6.2.1 Le type de données 63
6.2.2 La structure de classification 63
6.3 Partitionnement et critère de clustering 65
6.4 Méthodes de classification 66
6.4.1 Classement de méthodes de classification 66
6.4.2 Méthodes hiérarchiques de classification : les méthodes divisives et agglomeratives 67
6.5 Méthode de classification divisive pour les données symboliques 69
6.5.1 La matrice des données symboliques 69
6.5.2 Deux mesures de distances 71
6.6 L'extension du critère de variance intra-classes (within-class) . 73
6.7 Le bipartionnement d'un groupe 74
6.7.1 Question binaire et données symboliques 74
6.7.2 Choix de la meilleure bipartition 75
6.7.3 Choix du cluster à diviser 76
6.7.4 La règle d'arrêt et la sortie 76

7 Vue d'ensemble de l'analyse factorielle 78
7.1 Introduction 78
7.2 Définitions et hypothèses 78
7.3 Estimation des poids et des facteurs 81
7.3.1 Estimation et calcul des poids 81
7.3.2 Méthode en composantes principales 82
7.3.3 Estimation des facteurs 83
7.4 Les rotations 83
7.4.1 Les rotations orthogonales 84
7.4.2 les rotations obliques 84

II Recherche et résultats 85

8 Les fichiers 86
8.1 Les fichiers de données 86
8.1.1 Introduction 86
8.1.2 Le fichier *communes.xls* 87
8.1.3 Les fichiers *popact.xls, agelo.xls, tailmen.xls, popage98.xls, revenu.xls* 88
8.2 Les fichiers **.sds* 98

8.2.1 Introduction 98
8.2.2 Le programme DB2SO 99
8.2.3 La description des fichiers **.sds* 99
8.3 Les fichiers d'entrée et de sortie 100

9 Les résultats **101**
9.1 Recherches réalisées à partir du fichier *communes.xls* 101
9.2 Recherches réalisées à partir des fichiers *agelo.xls, popact.xls, revenu.xls, popage98.xls, tailmen.xls* 102
9.3 Interprétation des résultats du fichier *wal.resu* 103
9.3.1 Rappels : Division des classes 103
9.3.2 Première division 104
9.3.3 Deuxième division 105
9.3.4 Interprétations et critiques 106
9.3.5 Troisième division 111
9.3.6 Quatrième division 111
9.3.7 Interprétations et critiques 112
9.3.8 Cinquième division 118
9.3.9 Sixième division 118
9.3.10 Interprétation et critiques 118

10 Bugs **124**

Introduction

La littérature en matière de classification est depuis longtemps orientée vers le développement de nouvelles méthodes de classification. De plus, l'évolution constante et rapide des moyens informatiques durant ces dernières années permet à de nombreuses entreprises de recueillir de nombreuses informations et ainsi de traiter des fichiers de plus en plus gros. Par conséquent, il est important de trouver des méthodes qui gèrent ces très grosses bases de données et qui facilitent l'interprétation des résultats.
Nous nous intéresserons dans ce mémoire à la méthode monothétique divisive de classification symbolique réalisée dans le cadre du projet SODAS. Notre objectif est d'appliquer et d'évaluer les performances de cette méthode sur un ensemble de fichiers dont les variables décrivent les communes wallonnes.

Ce mémoire est divisé en deux grandes parties :

D'une part, nous exposerons la théorie nécessaire à la bonne compréhension de la méthode. Dans le premier chapitre, nous expliquerons dans quelles circonstances, cette méthode a été élaborée.
Dans le deuxième chapitre, nous rappellerons les méthodes classiques de statistique. Les deux chapitres suivants seront consacrés respectivement aux données et objets symboliques.
Le cinquième chapitre présentera les similarités et les dissimilarités correspondantes aux données symboliques.
Dans le sixième chapitre, nous parlerons des méthodes de classification symbolique. Enfin, le septième chapitre sera consacré à l'explication de l'analyse factorielle.

D'autre part, nous décrirons les différents fichiers nécessaires à l'utilisation de la méthode de classification et ensuite, nous exposerons et interpréterons les résultats obtenus. Enfin, nous signalerons quelques difficultés techniques.

Première partie

La partie théorique

Chapitre 1

Le projet SODAS (Système officiel de l'Analyse des Données symboliques)

1.1 Les objectifs

SODAS est un projet européen, supervisé par Eurostat.
Ce projet permet une utilisation plus facile des techniques de statistiques par les entreprises lorsque celles-ci doivent travailler avec de gros fichiers de données. L'objectif est d'extraire des données observées, qui peuvent être parfois de grande taille, une vue concise et structurée ainsi que des représentations facilement interprétables par l'utilisateur. Il s'agit également de proposer des outils mathématiques et informatiques permettant de modéliser et traiter des objets complexes i.e. des données structurées exprimant parfois une variation interne, et qui ne sont pas représentables naturellement par un point dans un espace euclidien.
Ce système permet aussi de communiquer plus facilement des données à des membres de la communauté de statistique et d'analyser des données qui ont une structure complexe. Ce projet est réalisé à un niveau européen pour homogénéiser les méthodes applicables aux objets symboliques et pour partager des concepts et des méthodologies sur des bases de données différentes.
Cela nécessite la participation de nombreux organismes de pays différents (FUNDP Belgique ; INRIA France ; DMS Italie ;...).

1.2 Le logiciel-prototype SODAS

Le résultat actuel du projet est un logiciel prototype qui comprend :

- Des outils génériques pour mémoriser, stocker et mettre sous forme de requête les objets symboliques qui formeront la structure de base des données pour représenter des données complexes.

- Une collection de méthodes d'analyse des données consacrées aux objets symboliques ; principalement, des méthodes descriptives univariées, méthodes de classification, construction d'arbre de décision, discrimination et analyse factorielle.

- Des facilités pour transformer les objets symboliques en des objets classiques et pour utiliser des méthodes classiques de l'Analyse des Données sur ceux-ci.

- Des outils ergonomiques pour présenter aux utilisateurs les résultats des méthodes employées.

1.3 Les méthodes d'Analyse des Données symboliques

Le but principal de l'approche symbolique dans l'Analyse des Données est d'étendre des problèmes, des méthodes et des algorithmes utilisés sur des données classiques à des données plus complexes où les unités sont appelées des objets symboliques. Ces objets apportent plus d'informations que les objets classiques au moins de deux façons : premièrement, dans le cas d'individus de complexité variable, en donnant la possibilité d'introduire une information structurée dans leur définition ; deuxièmement, dans le cas des classes, en étant défini "intentionnellement".

Les méthodes classiques de l'Analyse des Données prennent comme fichier input des tableaux à double entrées (INDIVIDUS, VARIABLES). Chaque cellule d'un tableau contient une valeur prise par une variable pour un indi-

vidu. Cette valeur est dite "atomique" dans le sens que ce n'est pas une liste ou un ensemble de mesures.

Par exemple :
Si les individus représentent les personnes d'un village et si les variables sont AGE et CATEGORIE SOCIO-PROFESSIONNEL (CSP) alors pour une personne, la cellule AGE contient une valeur (l'âge de la personne) et la cellule CSP contient une valeur (la CSP de la personne).

	AGE	CSP
Christelle	20	étudiante
John	32	agriculteur

Les méthodes de l'Analyse des Données symboliques introduites par le Professeur DIDAY prennent comme fichier input des tableaux à double entrées (INDIVIDUS,VARIABLES) où la valeur prise par un individu pour une variable peut être "non-atomique" i.e. un ensemble de valeurs, des intervalles des valeurs ou encore une distribution de probabilité.

Par exemple :
Si chaque "individu" représente un village (un ensemble de personnes) et si les variables sont toujours AGE et CSP alors une cellule de ce nouveau tableau contient pour chaque village, la liste des âges des personnes du village pour la variable AGE et la liste des CSP des personnes du même village pour la variable CSP.

	AGE	CSP
Merny	{15, 18, 28, 75}	{étudiant, agriculteur, pensionné}
Glaumont	{40, 45, 55}	{banquier, professeur}

1.4 Le plan de travail du projet

Le projet SODAS est divisé en 5 grandes parties :

- La construction, la gestion et la manipulation des objets symboliques.
- Les méthodes d'Analyse des Données symboliques.

- Lien avec l'Analyse des Données standards.
- Le rapport scientifique.
- L'utilisation des méthodes.

1.4.1 La construction, la gestion et la manipulation des objets symboliques

Le but de cette section est d'étudier, spécifier et développer des outils qui donnent des facilités pour traiter les objets symboliques. On cherche principalement à standardiser les descriptions des objets symboliques. Cette standardisation permettra une communication plus facile entre les différentes méthodes. Cette section a également pour tâche d'élaborer le logiciel SODAS final.

1.4.2 Les méthodes de l'Analyse des Données symboliques

L'objectif de cette partie est d'étendre les méthodes de l'Analyse des Données classiques à des objets symboliques. Les méthodes suivantes sont concernées : les méthodes de statistiques de base, les méthodes de classification, l'analyse discriminante et l'analyse factorielle.

Les méthodes de statistiques de base

Ce projet étend les méthodes de base des statistiques standards (moyenne, variance, covariance,...) à des exemples où les individus sont des objets symboliques. Des extensions d'histogrammes et des représentations d'autres graphiques sont également étudiées.

Les méthodes de classification

Les méthodes de classification symbolique aspirent à produire des groupes qui sont représentés par des objets symboliques. Ces méthodes ont l'avantage de proposer une classification avec une interprétation en termes de variables initiales.
De plus, elles permettent de prendre comme entrée des objets symboliques, qui sont des données plus complexes. Comme, c'est le cas pour des données

numériques, les méthodes de classification symbolique varient selon la structure recherchée. Parmi ces structures, les plus populaires sont les partitions et les arbres hiérarchiques. Les méthodes de classification non-paramétrique sont également étudiées.

Méthodes d'analyse discriminante sur les objets symboliques

Il existe deux axes de recherche pour les méthodes d'analyse discriminante :

Les Arbres de décision

Les arbres de décisions sont des outils efficaces pour construire des fonctions discriminantes dépendant de plusieurs variables. Ils sont aussi utilisés fréquemment dans la représentation des problèmes de classification. Cette section vise à étendre les techniques de segmentation binaire à l'analyse des données symboliques. Cela permet de prendre en compte des descriptions plus complexes et de manipuler des informations imprécises, partielles,...
Cet aspect est primordial car la plupart des techniques de partitionnement récursives sont très sensibles aux données incomplètes.

Les estimateurs de noyaux

Il est également proposé d'étendre les méthodes de l'analyse discriminante bayesienne non paramétrique aux cas des données symboliques. Une fois les classes obtenues, on peut assigner un nouvel objet à une classe en utilisant les estimateurs des noyaux.
Cela exige de connaître :

1. pour les objets probabilistes :
   - les densités estimées des classes
   - le calcul des probabilités pour les algorithmes
2. pour les objets intervalles et booléens :
   - le calcul des classes et de la généralisation
   - les règles de discrimination symbolique.

Analyse factorielle

Les méthodes d'analyse factorielle sont bien connues et largement utilisées pour réduire la dimension et détecter la structure essentielle de l'espace d'un ensemble de données.
Ces méthodes ne gèrent que des objets standards où chaque caractéristique a une valeur simple et précise. Ces méthodes sont donc également généralisées pour des données symboliques.

1.4.3 Lien avec l'analyse des données classiques

Une façon de traiter les objets symboliques consiste à transformer l'information "symbolique" en information numérique afin d'appliquer des procédures statistiques multivariées classiques.
L'analyse factorielle et les méthodes de classification sont des outils très puissants pour étudier des données multidimensionnelles. Mais, la complexité de telles données donnent un fichier de résultats très difficile à comprendre. Ce module vise à chercher une interprétation symbolique des résultats numériques multidimensionnels.
De plus, la plupart des bases nationales de statistiques sont "indicées" par le temps :

Par exemple :
Annuellement : le produit intérieur brut, le nombre d'étudiants dans une formation donnée.
Mensuellement : la consommation d'un produit donné, le nombre de personnes au chômage.
Journellement : le taux de change des monnaies étrangères,...

Ainsi, des objets symboliques sont utilisés pour interpréter les résultats de l'analyse des données temporelles en expliquant aussi clairement que possible les transformations temporelles.

Ce module requiert trois étapes :

1. L'élaboration de tables numériques classiques à partir des objets symboliques. Cette transformation exige la définition d'une distance entre des objets symboliques.
2. L'interprétation symbolique des classes et de l'analyse en composantes principales.
3. L'interprétation symbolique de l'analyse des données temporelles. De plus, la transformation d'une chaîne de tables de contingence en une chaîne de matrices de transition qui permettent de développer des assertions probabilistes avec des variables descriptives temporelles ordinales.

1.4.4 Le rapport scientifique

Cette quatrième partie a pour but l'élaboration d'un rapport scientifique unifié. Ce module est consacré aux tâches suivantes :

- le choix de notations adéquates
- la présentation et l'illustration des concepts théoriques de base
- la présentation unifiée des méthodes
- l'interaction entre les méthodes.

1.4.5 L'utilisation

Ce dernier module donne l'opportunité à des utilisateurs de tester les méthodes sur leurs propres données et également d'évaluer le logiciel-prototype et les interfaces. Ainsi, en cas de problème, des modifications pourront être apportées.

Chapitre 2

Rappels de méthodes classiques de statistiques

2.1 Introduction

Pour les méthodes classiques de l'Analyse des Données, les données sont obtenues pour des individus-singletons et chaque variable pour un individu ne contient qu'une seule valeur ou une seule catégorie.

Par exemple :
Le poids (variable) d'une personne (individu) est de 80 kg ; la marque (variable) d'une voiture (individu) est rouge,...

2.2 Les variables classiques

Considérons

- un ensemble de n individus $\Omega = \{1, ..., n\}$
- $Y_1, ..., Y_p$ les p caractéristiques, variables de chaque individu.
- $\mathcal{Y}_j$ est l'ensemble des observations où la variable Y_j prend ses valeurs $(j = 1, ..., p)$.

Définissons les variables :

$$\begin{aligned} Y_j : \Omega &\longrightarrow \mathcal{Y}_j \\ k &\longrightarrow Y_j(k) = x_{kj} \end{aligned}$$

où x_{kj} représente la valeur de la propriété j pour l'individu k.

Nous notons la **matrice des données** $\tilde{X} = (x_{kj})$ où $k \in \{1, ..., n\}$ et $j \in \{1, ..., p\}$.

Suivant la structure algébrique de $\mathcal{Y}_j$, il existe principalement deux types de variables :

- les variables quantitatives
- les variables qualitatives.

2.2.1 Les variables quantitatives

Une variable Y est dite **quantitative** lorsque l'ensemble des observations ou des valeurs possibles $\mathcal{Y}$ est inclus ou égal à $\mathbb{R}$.

Une variable quantitative est :

- **continue** lorsque $\mathcal{Y}$ est un intervalle continu dans $\mathbb{R}$.

  Par exemple : $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}_+$, $\mathcal{Y} = [2, 3]$.

- **discrète** lorsque $\mathcal{Y}$ contient un nombre fini ou infini dénombrable d'éléments de $\mathbb{R}$.

  Par exemple : $\mathcal{Y} = \{1, 2, 3, 4, ...\} = N_0$, $\mathcal{Y} = \{2, 4, 6, 9\}$.

Nous utilisons le plus souvent la distance euclidienne pour calculer la proximité entre deux éléments de $\mathcal{Y}$:

$$\forall x, y \in \mathcal{Y} : \delta(x, y) = \mid y - x \mid$$

2.2.2 Les variables qualitatives

Une variable Y est dite **qualitative** lorsque l'ensemble des observations est fini et contient des catégories qui n'ont aucune structure additionnelle et multiplicative.
Nous pouvons également distinguer les variables qualitatives nominales et ordinales.

Dans le cas de variables qualitatives **nominales**, l'ensemble des observations $\mathcal{Y}$ ne possède pas de structure interne, donc est non ordonné.

Exemples : $\mathcal{Y} = \{rouge, jaune, vert, bleu\}$, $\mathcal{Y} = \{homme, femme\}$ ou encore $\mathcal{Y} = \{0, 1\}$

Nous ne pouvons distinguer deux catégories $x, y \in \mathcal{Y}$ que de deux manières :

$$x = y \qquad ou \qquad x \neq y$$

les deux éléments sont soit égaux, soit différents.

La proximité entre deux catégories $x, y \in \mathcal{Y}$ est calculée par la formule suivante :

$$\delta(x, y) = \begin{cases} 1 & \text{si x} = \text{y} \\ 0 & \text{si x} \neq \text{y} \end{cases}$$

Remarque :
Lorsque l'ensemble des observations d'une variable Y ne comprend que deux éléments alors nous pouvons coder $\mathcal{Y} = \{homme, femme\}$ par $\mathcal{Y} = \{0, 1\}$
Une telle variable est appelée une **variable binaire.**
En fait, $Y(k) = 1$ signifie que l'individu k possède la propriété (donc est une femme),
$Y(k) = 0$ signifie que l'individu k ne vérifie pas la propriété (donc n'est pas une femme donc un homme).

D'autre part, les variables qualitatives sont **ordinales** lorsque l'ensemble des réalisations possibles est ordonné par l'ordre total $<$:

$$\forall \quad x, y \in \mathcal{Y} \quad : \quad x < y \quad ou \quad y < x$$

Exemple : Y = la qualité d'un produit avec $\mathcal{Y}$ ={excellent,bon, satisfaisant,insuffisant,mauvais}

La plupart du temps, les catégories de $\mathcal{Y}$ pour des variables ordinales sont encodées sous forme digitale :

$$\mathcal{Y} = \{0, 1, 2, 3, ..., s\}$$

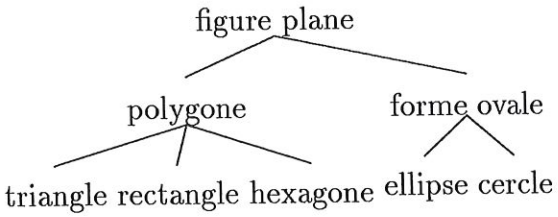
Les codes digitaux permettent de calculer une proximité entre deux éléments de $\mathcal{Y}$ qui sera considérée comme une échelle :

$$\forall \quad a, b \quad \in \mathcal{Y} \quad \delta(a, b) = \mid a - b \mid$$

Cas particuliers :

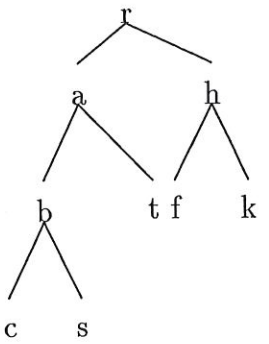
Les catégories de certaines variables ordinales sont ordonnées dans un arbre hiérarchique $\mathcal{H}$ qui est appelé **taxonomie**. De telles variables sont appelées des variables de **taxonomie**.

Exemple : Y = le type d'une figure à deux dimensions.



Terminologie sur la taxonomie

Soit Y une variable de taxonomie,
$\mathcal{Y} = \{a, b, ...z\}$ l'ensemble des observations
$\mathcal{H}$ = l'arbre hiérarchique.



1. Une catégorie **c** est un **descendant** de **a** ou **a** un **ascendant** de **c** si **c** < **a**
2. **b** est un **successeur** ou un **descendant direct** de **a** si

$$\mathbf{b} < \mathbf{a} \quad et \quad \nexists\ \mathbf{d} \in \mathcal{Y} \quad tq \quad \mathbf{b} < \mathbf{d} < \mathbf{a}$$

3. **a** est l'**ascendant direct** ou le **prédecesseur** de **b** si

$$\mathbf{b} < \mathbf{a} \quad et \quad \nexists\ \mathbf{e} \in \mathcal{Y} \quad tq \quad \mathbf{b} < \mathbf{e} < \mathbf{a}$$

4. Chaque catégorie $\mathbf{d} \in \mathcal{Y}$ est un **noeud** de l'arbre hiérarchique $\mathcal{H}$
5. L'arbre hiérarchique $\mathcal{H}$ ne possède qu'une seule **racine** : $\exists!\ r \in \mathcal{Y}$ qui ne possède aucun prédécesseur.
6. Une catégorie qui n'a pas de successeur est appelée une **feuille** f de l'arbre ou un **noeud terminal**.

Il n'existe pas de proximité $\delta(x, y)$ entre deux noeuds x, y de $\mathcal{H}$ mais on parlera plutôt du chemin effectué entre les noeuds dans l'arbre $\mathcal{H}$.

2.3 La matrice des données

Considérons un ensemble de n individus $\Omega = \{1, ..., n\}$ et p variables $Y_1, ..., Y_p$.
Les variables Y_j prennent respectivement leurs valeurs dans les ensembles d'observations $\mathcal{Y}_j \quad j = 1, ..., p$.
Désignons par X le vecteur des p variables $Y_1, ..., Y_p$ qui prend ses valeurs dans χ :

$$X = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} \in \chi = \bigotimes_{j=1}^{p} \mathcal{Y}_j$$

où χ est le produit cartésien des différents ensembles d'observations $\mathcal{Y}_j$ correspondants aux variables Y_j $(j = 1, ..., p)$.

Comme nous l'avons déjà vu ultérieurement, nous notons $Y_j(k) = x_{kj}$:

$$\begin{array}{rccl} \forall k \in \Omega & Y_j : \Omega & \longrightarrow & \mathcal{Y}_j \\ & k & \longrightarrow & Y_j(k) = x_{kj} \end{array}$$

où x_{kj} est la valeur ou la catégorie observée de la variable Y_j pour l'individu k

Si on considère le vecteur colonne X, alors on obtient :

$$X(k) = (Y_1(k), ..., Y_p(k))' = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kp} \end{pmatrix} = x_k \qquad k = 1, ..., n$$

où le vecteur colonne $x_k \in \chi$ donne les valeurs prises par chaque variable pour l'individu k.

La matrice des données classiques est une matrice à n lignes et p colonnes :

$$\tilde{X} = (x_{kj})_{n\times p} = \begin{pmatrix} x_1' \\ \vdots \\ x_k' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

où la ligne x_k' représente les caractéristiques obtenues pour l'individu k.

Nous pouvons noter la matrice des données sous la forme suivante :

$$\tilde{X} = \begin{pmatrix} y_1 & \cdots & y_j & \cdots & y_p \end{pmatrix}$$

où le vecteur y_j donne les valeurs prises par la variable Y_j pour tous les individus.

2.4 Les données manquantes

En pratique, les données sont amassées à partir d'interview, de questionnaires,...
Dans la plupart des cas, les matrices des données sont incomplètes.
Les valeurs ne sont pas disponibles pour diverses raisons :

- une personne ne répond pas à la question
- la personne sélectionnée est absente
- le coût est trop grand pour obtenir la valeur d'une variable.

Dans la matrice des données, nous affecterons dans la cellule correspondante à la valeur manquante le sigle * ou le mot NULL.
Les valeurs manquantes posent un sérieux problème, elles introduisent un résultat biaisé dans l'analyse de la matrice des données.

Plusieurs procédés ont été introduits pour minimiser ce problème :

1. ne considérer que les variables Y_j pour lesquelles aucune valeur ne manque dans la matrice des données
2. remplacer chaque valeur manquante x_{kj} par une valeur plausible à partir des autres données.

   Par exemple : Pour des variables qualitatives, nous pouvons remplacer la valeur manquante par la moyenne des autres valeurs présentes : $\overline{y}_j = \sum_{k=1}^{n} x_{kj}$.

2.5 Exemple

Soit Ω ={Christelle,Anne,Thérèse,Pierre,Samuel,Laurent} un ensemble de six étudiants pour lequel les quatre variables suivantes ont été considérées :

- Y_1 = la taille en cm
- Y_2 = le sexe (0=femme ou 1=homme)
- Y_3 = le grade obtenu (A,B,C,D,E)
- Y_4 = la nationalité (f=français, l=luxembourgeois, b=belge, h=hollandais).

Les variables Y_1, Y_2, Y_3, Y_4 sont des variables respectivement quantitative, qualitative nominale (binaire), qualitative ordinale, nominale (avec 4 categories).
La matrice des données $\tilde{X}$ correspondante où le nombre de variables p=4 et le nombre d'individus n=6 est de la forme suivante :

	Y_1	Y_2	Y_3	Y_4
Christelle	*	0	A	b
Laurent	175	1	C	l
Thérèse	164	0	B	l
Pierre	185	1	A	b
Samuel	180	1	E	f
Anne	156	0	B	h

Nous pouvons remarquer que chaque cellule de la matrice ne contient qu'un seul élément (singleton).

Chapitre 3

Les données symboliques

3.1 Introduction

Nous avons décrit au chapitre précédent la matrice des données classiques $\tilde{X} = (x_{kj})_{n \times p}$ pour l'ensemble $\Omega = \{1, ..., n\}$ et chaque cellule de cette matrice ne contenait qu'une seule valeur ou catégorie.
L'arrivée de systèmes de saisie automatique d'informations dans tous les secteurs d'activités provoque une accumulation de données. Extraire de celles-ci des informations utiles reste un problème majeur pour les entreprises. Mais l' évolution technologique en informatique et les progrès dans l'Analyse des Données résolvent partiellement la complexité grandissante des données. Sur la base des travaux du projet SODAS, nous introduisons la notion de données symboliques.

3.2 Préliminaires

Il existe trois types de données symboliques qui seront définies dans la suite :

- variables à valeurs multiples
- variables intervalles
- variables modales.

Définissons l'ensemble des "objets" E de deux façons différentes :

1. l'univers $E = \Omega = \{1, ..., n\}$ est un ensemble d'individus élémentaires appelés des **objets du premier ordre.**
2. Un système $E = \{C_1, C_2, ...\}$ est un ensemble de classes $C_i \subseteq \Omega$ appelées **des objets du second ordre.**

Remarque :
Dans certaines situations, l'ensemble Ω de n individus est un échantillon d'une population beaucoup plus grande. Par exemple, on peut prendre au hasard n villes dans un pays ou n voitures d'une ligne de production.

3.3 Variables définies dans un ensemble de valeurs

3.3.1 Définitions

Soit $\mathcal{Y}$ = l'ensemble des observations possibles.
Une variable Y est appelée une **variable à valeur dans un ensemble** $\mathcal{B}$:

$$\begin{array}{rccc} \forall k \in E \quad Y: & E & \longrightarrow & \mathcal{B} \\ & k & \rightsquigarrow & Y(k) \end{array}$$

où $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset \mid U \subseteq \mathcal{Y}\}$.

Considérons deux types de variables à valeur dans un ensemble :

- les variables à valeurs multiples
- les variables intervalles.

1. Une variable Y est dite **à valeurs multiples** lorsque les valeurs $Y(k)$ sont toutes des sous-ensembles finis de $\mathcal{Y}$:

$$| Y(k) | < \infty$$

Remarque :
Si les éléments de $\mathcal{Y}$ sont des catégories alors nous parlerons de variables **à catégories multiples.**
Si les éléments de $\mathcal{Y}$ sont des nombres rééls i.e. $Y(k) \subset I\!R$ alors nous parlerons de variables **à valeurs quantitatives multiples.**

Exemple :
Soit $E = \Omega$={les capitales d'Europe}

- Y_1 représente les banques présentes dans une capitale telle que $\mathcal{Y}$ ={BNP, CL, DB, Dres, SPk, Lux, BdiRoma, Barc, CL, DB} est l'ensemble des banques possibles.
- Y_2 représente les taxes payées par les deux plus grandes entreprises (en euro) où $\mathcal{Y} = I\!R^+$ est l'ensemble des solutions possibles.

Voici les résultats pour 5 individus de l'ensemble E :

	Y_1	Y_2
Paris	{BNP,CL,DB,Dres}	{900.000,750.000}
Bonn	{SPk,DB,Dres,Lux}	{1.200.000,1.050.000}
Rome	{DB,BdiRoma}	{800.000,650.000}
Luxembourg	$\mathcal{Y}$	{1.150.000,1.100.000}
Londres	{Barc,CL,DB}	{850.000,850.000}

Y_1 est une variable à catégories multiples.
Y_2 est une variable quantitative à valeurs multiples.

2. Une variable Y est appelée **une variable intervalle** si :

$$\begin{array}{rccl} Y & : & E & \longrightarrow & \mathcal{B} \\ & & k & \rightsquigarrow & Y(k) = U \end{array}$$

où

- U est un intervalle de $I\!R$ (intervalle fermé ou non)
- $\mathcal{B}$ est l'ensemble de tous les intervalles $\mathcal{I}$ de $\mathcal{Y}$

Exemple :

- E = { les étudiants de deuxième licence en sciences mathématiques }.
- Y = le temps pour venir à l'université (en minutes).
- $\mathcal{Y} = \{[a,b] \mid a,b \in I\!R^+, 0 \leq a \leq b < \infty\}$

	Y
Cristelle	[20,25]
Laurent	[6,10]
Thérèse	[15,18]
Anne	[10,13]
Pierre	[11,16]
Samuel	[14,18]

Les exemples ci-dessus considérent l'ensemble $E = \Omega$ c'est-à-dire l'ensemble d'individus élémentaires (objets du premier ordre).

Maintenant, supposons que

- $\Omega = \{1, ...n\}$ sont des objets du premier ordre.
- $\tilde{Y}$ est une variable à une seule valeur.
- $E = \{C_1, ..., C_m\}$ est l'ensemble de classes $C_i \subseteq \Omega$ qui sont appelées des objets du second ordre.

Nous cherchons à caractériser le comportement de ces classes par rapport à la variable $\tilde{Y}$.
L'approche proposée par le projet Sodas est de définir une variable Y "globale".

Voici deux illustrations pour mieux comprendre cette approche :

1. Considérons
   - $E = \{C_1, ..., C_m\}$ l'ensemble des m classes C_i d'une école.
   - $\Omega = \sum_{i=1}^{m} C_i$ l'ensemble de tous les élèves.
   - $\tilde{Y}(k)$ représente la hauteur d'un étudiant en mètres (une variable classique à une seule valeur).

- $Y(C_i)$ est l'ensemble des hauteurs trouvées dans la classe C_i

Supposons que dans la classe C_1, nous avons quatre élèves et que $Y(C_1)$ $=\{1.50,1.56,1.73,1.80\}$.
Nous avons transformé la variable $\tilde{Y}$ classique à valeur simple en une variable quantitative à valeurs multiples.

2. $Y(C_i)$ donne un intervalle de borne inférieure égale à la plus petite hauteur d'un élève la classe et de borne supérieure égale à la plus grande hauteur de la classe C_i :

$$Y = [min, max] = [\alpha, \beta] \text{ est un intervalle de } \mathcal{Y} = I\!R^+$$

avec

$$\alpha = min_{w \in C_i}\{\tilde{Y}(w)\}$$

$$\beta = max_{w \in C_i}\{\tilde{Y}(w)\}$$

et par conséquent, pour les quatre élèves de la classe C_1 :

$$Y(C_1) = [1.50, 1.80]$$

Dans ce cas, nous obtenons à partir de la variable $\tilde{Y}$ une variable quantitative-intervalle.

En résumé, nous avons défini une variable globale à valeurs dans les sous-ensembles $Y(C) \subseteq \mathcal{Y}$ pour tous les sous-ensembles $C \subseteq P(\Omega)$

3.4 Variable modale

3.4.1 Préliminaires

Dans la plupart des cas, une variable modale Y définie sur l'ensemble des objets $E = \{a, b, ...\}$ est une variable à états multiples où pour chaque objet $a \in E$, nous obtenons la catégorie $Y(a) = y \subseteq \mathcal{Y}$ et une fréquence, une probabilité ou un poids $w(y)$ qui indique la pertinence de la catégorie y pour l'objet a.

Illustrons ce concept par l'exemple suivant :

- $\Omega = \{1, ..., n\}$ un ensemble de n individus.
- $\tilde{Y}$ est une variable-catégorie à valeur simple.
- $E = P(\Omega)$ est l'ensemble de classes d'individus $C \neq \emptyset$.

Considérons l'ensemble des catégories de $\mathcal{Y}$ qui sont observées dans la classe C :

$$U(C) = \{\tilde{Y}(k) \mid k \in C\} \subseteq \mathcal{Y}$$

Pour chaque catégorie $y \in U(C)$, la fréquence relative $w_c(y)$ de la catégorie y dans la classe C :

$$w_c(y) = \frac{\mid \{k \in C \mid \tilde{Y}(k) = y\} \mid}{\mid C \mid}$$

où $w_c(y)$ sont appelés les poids.

Nous définissons par conséquent la variable modale globale :

$$Y(C) = (U(C), w_c) \quad \forall C \in E$$

3.4.2 Définitions

Une variable **modale** Y sur un ensemble $E = \{a, b, ...\}$ d'objets à valeurs dans $\mathcal{Y}$ est :

$$Y(a) = (U(a), \pi_a) \quad \forall a \in E$$

où

- π_a est une mesure non négative ou une distribution (fréquence, probabilité ou poids) sur les valeurs possibles de $\mathcal{Y}$.
- $U(a) \subseteq \mathcal{Y}$ est le support de π_a dans le domaine $\mathcal{Y}$.

Remarque :
Dans certains cas, nous ne sommes pas obligés de parler du support de U(a) puisque il est déterminé implicitement par π_a et donc $Y(a) = \pi_a \ \ \forall a \in E$:

$$Y : E \longrightarrow \mathcal{B} = \mathcal{M}(\mathcal{Y})$$

où $\mathcal{M}(\mathcal{Y})$ est la famille des mesures non-négatives sur $\mathcal{Y}$ tel que $Y(a) = \pi_a$. Une variable modale assigne une mesure ou une probabilité à chaque objet $a \in E$.

L'interprétation de ces poids dépend principalement de l'application sous-jacente :

1. Dans une étude de marketing pour un certain nombre de produits (catégories) $y \in \mathcal{Y}$, $\pi_a(y) = \pi_a(\{y\})$ peut représenter le nombre de fois qu'un consommateur a a acheté le produit y l'année dernière.

2. Dans une interprétation fréquentiste, en considrant a comme une classe d'individus C, $\pi_C(B)$ est, pour chaque sous-ensemble B, la fréquence avec laquelle les valeurs ou les catégories d'un ensemble B sont observées dans la classe C.

   Par exemple :
   Si l'individu k représente la Belgique alors la variable Y "langue linguistique" peut être spécifiée par une distribution de fréquence :

   $$Y(k) = ((Neerlandais, 0.50), (Francais, 0.45), (Allemand, 0.05))$$

   Ce qui signifie entre autre que 50% des belges parlent le Néerlandais (les langues non-citées ont donc une fréquence nulle).
   Cette distribution pour le pays k (i.e. la Belgique) peut être représentée par un diagramme en barres.

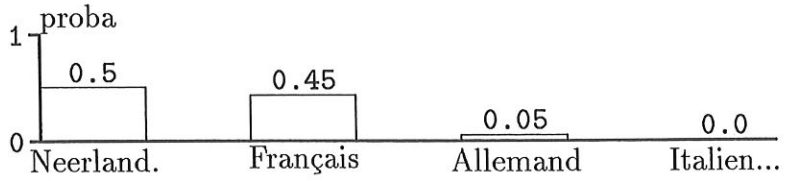


Figure : *distribution des langues officielles en Belgique*

Remarque :
Les variables qui permettent une représentation des valeurs par un diagramme en barres seront appelées **des variables diagrammes.**
Y est une **variable diagramme** si $\mathcal{Y}$ est fini et π_a peut être représenté par un diagramme en barres.
Y est **une variable histogramme** si π_a peut être spécifié par un histogramme.

3. Dans un sens probabiliste et avec un individu élémentaire a, $\pi_a(B)$ peut désigner la probabilité que l'individu a prenne sa valeur dans l'ensemble B.

3.5 Le tableau des données symboliques

Soit E un ensemble de N entités $E = \{1, ..., N\}$,
les entités u de E sont appelées des objets et :

- soit $E{=}\Omega = \{1, ..., n\}$ est un ensemble de n individus k $(N = n)$
- soit E est un sous-ensemble de Ω i.e. $E \subseteq \Omega$ $(N \leq n)$
- soit E est une collection de classes d'individus $C_1, ..., C_m \subseteq \Omega$ $(N = m$, les objets sont du second ordre).

Supposons que chaque entité $u \in E$ soit décrite par p variables symboliques $Y_1, ..., Y_p$ respectivement à valeurs dans $\mathcal{Y}_1, ..., \mathcal{Y}_p$:

$$\forall j \in 1, ..., p : \quad Y_j : E \longrightarrow \mathcal{B}_j$$

où

- dans le cas classique : $\mathcal{B}_j \in I\!R$
- $\mathcal{B}_j = \mathcal{I} = \{[\alpha, \beta] \mid -\infty < \alpha \leq \beta < \infty\}$ est un ensemble d'intervalles.
- un ensemble de catégories nominales,ordinales,... tel que

$$\mathcal{B}_j = \{B \mid B \subseteq \{a, b, ..., t\}\}$$

- $\mathcal{B}_j = \mathcal{M}(\mathcal{Y}_j)$ est la famille des mesures non négatives (poids, fréquences, probabilités) sur $\mathcal{Y}_j$

Notons $X(u) = (Y_1(u), ..., Y_p(u))'$ le vecteur des variables symboliques pour l'objet $u \in E$.
Chaque entité $u \in E$ peut être décrite par le vecteur de données symboliques $x_u = X(u) = (\xi_{u1}, ..., \xi_{up})'$ où $\xi_{uj} = Y_j(u) \in \mathcal{B}_j$ est le vecteur des valeurs prises par les variables symboliques Y_j pour l'entité u $(j = 1..., p)$.

Définissons la **matrice des données symboliques** :

$$\underline{X} = \begin{pmatrix} x_1' \\ \vdots \\ x_N' \end{pmatrix} = (\xi_{uj})_{N \times p} = \begin{pmatrix} \xi_{11} & \cdots & \xi_{1p} \\ \vdots & \ddots & \vdots \\ \xi_{N1} & \cdots & \xi_{Np} \end{pmatrix}$$

où

- x_u' est le vecteur des données symboliques pour l'entité u qui est appelée le **vecteur description** de l'objet $u \in E$.
- ξ_{uj} est la valeur de la variable j pour l'entité u, mais cette "valeur" peut être un intervalle, un histogramme, un ensemble de valeurs,...
  Les ξ_{uj} peuvent être de types différents (j=1,...p).

Exemple 1 :
soit E={New-york, Washington, Los Angeles, Boston} est un ensemble de 4 villes $(N = 4)$.
Considérons 3 variables symboliques Y_1, Y_2, Y_3 :

1. Y_1 = le nombre d'habitants en millions d'habitants (10^6) pour l'année 1998 avec les "valeurs" données sous la forme d'intervalle :

   $[\alpha, \beta] = [min, max]$ avec $0 < \alpha \leq \beta < \infty$ et $\alpha, \beta \in I\!R$

   où
   - α est le nombre minimum d'habitants en 1998
   - β est le nombre maximum d'habitants en 1998
   - $\mapsto Y_1$ est une variable intervalle quantitative.
2. Y_2 représente le pourcentage obtenu par les partis politiques à la dernière élection
   où

- $\mathcal{Y}_2$={**D**émocrate, **C**onservateur, **S**ocialiste, **N**ationaliste}
- $\mathcal{B}_2$ est l'ensemble des fréquences sur $\mathcal{Y}_2$
- $\mapsto Y_2$ est une variable modale et une variable histogramme.

3. Y_3 représente la liste des banques représentées dans une ville où
   - $\mathcal{Y}_3$ ={CL,SPK,DB,BNP,...}
   - $\mathcal{B}_3$ est un sous-ensemble de $\mathcal{Y}_3 = \mathcal{P}(\mathcal{Y}_3)$
   - $\mapsto Y_3$ est une variable multi-états.

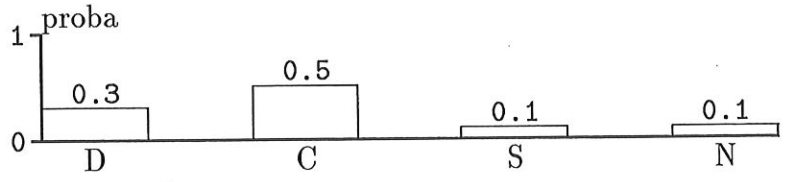
On obtient une matrice des données symboliques $\underline{X}$ à 4 lignes et à 3 colonnes :

	Y_1	Y_2	Y_3
New-York	(80,100]	(D (0.4) ; C (0.3) ; S (0.2) ; N (0.1))	{CL,BNP}
Washington	(100,130]	(D (0.1) ; C (0.3) ; S (0.4) ; N (0.2))	{SPK,DB,BNP}
Los Angeles	(8,10]	(D (0.3) ; C (0.5) ; S (0.1) ; N (0.1))	{SPK}
Boston	(10,13]	(D (0.3) ; C (0.3) ; S (0.3) ; N (0.1))	{CL,DB,BNP}

Examinons le vecteur description de la ville de Los Angeles :

$$x_3' = ((8,10], \quad (D(0.3); C(0.5); S(0.1); N(0.1)), \quad \{SPK\})$$

Ce vecteur peut être représenté d'une manière différente en représentant les probabilités obtenues pour la variable Y_2 en un diagramme en barres :



Exemple 2 :
Cet exemple montre qu'il est possible d'obtenir une matrice des données symboliques à partir de la matrice des données classiques, i.e de transformer des variables classiques $\tilde{Y}$ en variables globales symboliques Y.

Rappels :

- Ω est l'ensemble de n individus (objets du premier ordre).
  Les n individus sont décrits par p variables classiques $\tilde{Y}_1, ..., \tilde{Y}_p$.

- E est une collection de classes $= \{C_1, ..., C_m\}$ telles que $C_i \subseteq \Omega \quad \forall i = 1, ..., m$ (des objets du second ordre).

Transformons les variables classiques en variables globales :

$$Y_j(C_i) = \{\tilde{Y}_j(k) \mid k \in C_i\} \subseteq \mathcal{Y}_j$$

l'ensemble des valeurs qui sont observées pour $\tilde{Y}_j$ pour les éléments de C_i.

Un vendeur de voiture a un stock de 15 voitures d'occasion qui sont décrites par 3 variables :

- $\tilde{Y}_1$ représente la marque d'une voiture
- $\tilde{Y}_2$ donne le prix d'une voiture en euro (en milliers d'euro)
- $\tilde{Y}_3$ représente le pays de construction de la voiture.

La matrice (15×3) des données classiques $\tilde{X}$ est représentée par la table suivante :

	$\tilde{Y}_1$	$\tilde{Y}_2$	$\tilde{Y}_3$
1	citroen	22	France
2	citroen	17	France
3	peugeot	20	France
4	peugeot	15	France
5	renault	20	France
6	audi	21	Allemagne
7	mercedes	20	Allemagne
8	mercedes	25	Allemagne
9	bmw	23	Allemagne
10	bmw	24	Allemagne
11	bmw	30	Allemagne
12	fiat	16	Italie
13	fiat	12	Italie
14	fiat	14	Italie
15	alpha romeo	20	Italie

Si on considère les groupes :

- $C_1 = \{1, 2, 3, 4, 5\}$ classe des voitures françaises.
- $C_2 = \{6, 7, 8, 9, 10, 11\}$ classe des voitures allemandes.
- $C_3 = \{12, 13, 14, 15\}$ classe des voitures italiennes.

Nous obtenons la matrice des données symboliques $\underline{X}$ donnée par :

	Y_1	Y_2	Y_3
C_1	{citroen,peugeot,renault}	[15,22]	France
C_2	{audi, mercedes, bmw}	[20,30]	Allemagne
C_3	{fiat,alpha roméo}	[12,20]	Italie

avec trois variables symboliques :

- Y_1 = l'ensemble des marques de voiture présentes dans une classe
- Y_2 = variable intervalle ayant comme borne inférieure et supérieure respectivement les prix minimum et maximum présents dans une classe.
- Y_3 = le pays de construction d'un groupe.

Remarque :
Si on avait pris la classe C={2,4,12,13,14} des voitures les moins chères (avec $\tilde{Y}_3 < 20$) ; le vecteur description de ce groupe deviendrait :

$$((citroen\ 1; peugeot\ 1; fiat\ 3),\ [12, 17],\ \{France, Italie\})$$

où Y_1 serait une variable modale.

Chapitre 4

Les objets symboliques

4.1 Introduction

Dans le chapitre précédent, nous avons présenté différents types de données symboliques. Nous rappelons que les types de données symboliques sont des généralisation des types de données classiques.

L'ensemble des objets E est égal :

- soit à Ω l'ensemble des individus, objets élémentaires.
- soit à l'ensemble des classes d'individus $\{C_1, ..., C_m\}$.

Les propriétés de chaque objet $u \in E$ sont décrites par des variables symboliques notées $Y_1, ..., Y_p$ qui peuvent être des variables à valeurs multiples, des variables intervalles et des variables modales.
Il serait naturel d'appeler un objet $u \in E$ (qui est décrit par des variables symboliques) un objet symbolique ou un objet symbolique réel.
Cependant, le terme "objet symbolique" est utilisé dans un sens plus général.

Le but de l'analyse des données symboliques est d'examiner, visualiser, classer et réduire l'information qui est contenue dans la matrice des données symboliques $\underline{X} = (\xi_{uj})_{N \times p}$ pour les objets de E.
Le problème de base est la détection de tous les objets $u \in E$ stockés dans une base de données qui satisfont à des exigences apportées par des variables symboliques. Autrement dit, l'objectif est d'extraire de la table des données les individus ou objets correpondants à certaines exigences.
Ces exigences sont rassemblées dans une "query" (une assertion) qui est

identifiée par un objet symbolique **virtuel**.
Cet objet symbolique virtuel est une spécification formelle des propriétés des objets réels qui nous intéressent.

Exemple :
Prenons le tableau suivant :

	Y_1	Y_2	Y_3	Y_4
Christelle	163	0	A	b
Laurent	175	1	C	l
Therese	164	0	B	l
Pierre	185	1	A	b
Samuel	180	1	E	f
Anne	156	0	B	h

Une ligne du tableau caractérise un individu ou un objet.
Elle peut s'exprimer sous la forme d'une conjonction de propriétés.
Par exemple pour Pierre, ce quatrième individu est caractérisé par la conjonction suivante :

$$[Y_1 = 185] \bigwedge [Y_2 = 1] \bigwedge [Y_3 = A] \bigwedge [Y_4 = b]$$

Et si nous voulons extraire les individus qui vérifient la "query" suivante :

$$q = [Y_1 \subseteq [170, 190] \bigwedge [Y_3 \in \{A, E\}]$$

alors les individus vérifiant cette condition seront Pierre et Samuel.

Nous allons introduire quatre types d'objets symboliques :

- les objets-assertion
- les objets symboliques individuels
- les objets symboliques booléens
- les objets symboliques modaux.

4.2 Relations et descriptions

4.2.1 Introduction

Considérons $E = \{1, ..., N\}$ comme un ensemble d'objets (individus ou classes) décrits par p variables symboliques $Y_1, ..., Y_p$ définies respectivement dans les ensembles $\mathcal{Y}_1, ..., \mathcal{Y}_p$ et qui prennent respectivement leurs valeurs dans $\mathcal{B}_1, ..., \mathcal{B}_p$. Examinons les objets u de E pour lesquels les variables $Y_j(u)$ des vecteurs des données symboliques $Y(u) = (Y_1(u), ..., Y_p(u))'$ remplissent une ou plusieurs conditions :

Exemples de conditions possibles :

1. pour des variables Y_j classiques à une seule valeur :

$$[Y_j(u) \leq z_j], \quad [Y_j(u) \in \{rouge, bleu\}]$$

où $z_j \in \mathcal{Y}_j$ est un seuil

2. pour les variables Y_j symboliques (modales,intervalles,...) :

$$[Y_j(u) \subseteq D_j], \qquad \{Y_j(u) \subseteq [4, 6]\}$$

où D_j est un sous-ensemble spécfique de $\mathcal{Y}_j$

Pour construire des conditions, on peut utiliser de nombreuses relations $\mathcal{R}$. Par exemples, $=$, $\neq$, $\leq$, $\subseteq$, ...

4.2.2 Rappelons la terminologie utilisée pour les relations

Définitions

Soit deux ensembles W, Z
Une **relation** $\mathcal{R}$ définie sur le produit cartésien de $W \times Z$ est une fonction binaire :

$$\begin{aligned} \phi : W \times Z &\longrightarrow [0, 1] \\ (w, z) &\longrightarrow \phi(w, z) = [w\mathcal{R}z] = \begin{cases} 1 & \text{si } \mathcal{R} \text{ vraie pour } (w, z) \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

Quelques relations que nous pouvons rencontrer :

1. Une relation $\mathcal{R}_j$ définie sur $\mathcal{Y}_j \times \mathcal{Y}_j$ et donc de type $y_j \mathcal{R}_j z_j$ avec $\mathcal{R}_j \in \{\leq, >, =, ...\}$.

   Exemples : $Y_j(u) \leq 5, \quad Y_j(u) \geq 1.3$

2. Une relation $\mathcal{R}_j$ définie sur $\mathcal{Y}_j \times \mathcal{P}(\mathcal{Y}_j)$ et donc de type $y_j \mathcal{R}_j D_j$ avec $\mathcal{R}_j \in \{\in\}$.

   Exemple : $Y_j(u) \in \{rouge, bleu\}$

3. Une relation $\mathcal{R}_j$ définie sur $\mathcal{P}(\mathcal{Y}_j) \times \mathcal{P}(\mathcal{Y}_j)$ et donc de type $D'_j \mathcal{R}_j D_j$ avec $\mathcal{R}_j \in \{\subseteq, \supset, ...\}$.

   Exemples : $Y_j(u) \subseteq \{rouge, bleu\}, \; Y_j(u) \subseteq [4, 6]$

4. Une relation $\mathcal{R}_j$ est de type $y_j \mathcal{R}_j D_j$ entre une mesure $y_j \in \mathcal{M}(\mathcal{Y}_j)$ (ensemble des poids possibles sur $\mathcal{Y}_j$) et un ensemble $D_j \subseteq \mathcal{Y}_j$ c'est-à-dire définie sur $\mathcal{M}(\mathcal{Y}_j) \times \mathcal{P}(\mathcal{Y}_j)$.

   Exemple : $Y_j(u)(D_j) \geq 0.5$
   où
   - $y_j = Y_j(u)$
   - $\mathcal{R}_j =$ "$poids \geq 0.5$"

Définitions

1. Soit une **collection** $(\mathcal{R}_1, ..., \mathcal{R}_p)$ **de relations** où $\mathcal{R}_j$ est définie sur le produit cartésien $W_j \times Z_j$ tel que $w_j \mathcal{R}_j z_j$, $w_j \in W_j$ et $z_j \in Z_j$.
   Soit les produits cartésiens $W = \bigotimes_{j=1}^{p} W_j$ et $Z = \bigotimes_{j=1}^{p} Z_j$ comprenant respectivement
   - $w = (w_1, ..., w_p)'$
   - $z = (z_1, ..., z_p)'$.

   Alors la **relation produit** $\mathcal{R} = \bigotimes_{j=1}^{p} \mathcal{R}_j$ est définie sur $W \times Z$ par :

$$[w \mathcal{R} z] = \bigwedge_{j=1}^{p} [w_j \mathcal{R}_j z_j] = [w_1 \mathcal{R}_1 z_1] \bigwedge ..... \bigwedge [w_p \mathcal{R}_p z_p]$$

Les relations sont utilisées pour établir les "queries" q.
En fait, les "queries" sont des combinaisons de relations de type $[Y_j \mathcal{R}_j z_j]$ et $[Y_j \mathcal{R}_j D_j]$.

Exemple :
$q = [Y_1 \leq z_1] \bigwedge [Y_2 > z_2] \bigwedge ..... \bigwedge [\mid Y_p - z_p \mid \leq 0.1] = [Y \mathcal{R} z]$
$q = [Y_1 \in D_1] \bigwedge [Y_2 \supset D_2] \bigwedge ..... \bigwedge [Y_p(D_p) \geq 0.5] = [Y \mathcal{R} D]$

2. Soit $\mathcal{Y}_j$ l'ensemble des observations de la variable Y_j ; $j = 1, ..., p$.
Chaque élément $z = (z_1, ..., z_p) \in \chi = \bigotimes_{j=1}^{p} \mathcal{Y}_j$ est appelé **un vecteur description.**
Chaque p-uplets $(D_1, ..., D_p)$ d'ensembles $D_i \subseteq \mathcal{Y}_j$ est appelé **un système description.**
Chaque sous-ensemble $D \subseteq \chi$ est **un ensemble description.**
Un **ensemble description cartésien** D est un ensemble description $D = D_1 \times .... \times D_p$ construit à partir d'un système description $(D_1, ...D_p)$.

Exemple :
Soit trois variables classiques quantitatives $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$ définies sur Ω et à valeurs respectivement dans $\mathcal{Y}_j = I\!R, \quad \forall j = 1, ..., 3$.
Alors, le vecteur z=(100,20,7) $\in \chi = \bigotimes_{j=1}^{3} \mathcal{Y}_j = I\!R^3$ est un vecteur description et par exemple, le produit cartésien $D \subseteq \chi = I\!R^3$ est :

$$D = [0, 100] \times (10, 20) \times (4, 7]$$

est un ensemble description. Intuitivement, D correspond à une "query" q telle que :

$$q = [\tilde{Y}_1 \in [0, 100] \bigwedge \tilde{Y}_2 \in (10, 20) \bigwedge \tilde{Y}_3 \in (4, 7] \ ]$$

4.3 Objets assertions

4.3.1 Définitions

Intuitivement, un objet assertion est une "conjonction de conditions".
Soit un vecteur $Y = (Y_1, ..., Y_p)'$ de variables symboliques ou classiques Y_j

définies sur E et à valeurs dans $\mathcal{Y}_j$, $j = 1, ..., p$.
Soit des éléments fixes $z_j \in \mathcal{Y}_j$.

1. Une condition du type $[Y_j \mathcal{R}_j z_j]$ est appelée un **évènement élémentaire** si la fonction binaire :

$$\begin{array}{rcl} \phi : E & \longrightarrow & \{0,1\} \\ u & \longrightarrow & \phi(u) = [Y_j(u)\mathcal{R}_j z_j] = \left\{ \begin{array}{ll} 1 & \text{si la relation } Y_j(u)\mathcal{R}_j z_j \text{ est vrai} \\ 0 & \text{sinon} \end{array} \right. \end{array}$$

2. Prenons r indices au hasard dans l'ensemble $\{1, ..., p\}$ des indices des p variables notées $j_1, ..., j_r$ $(r < p)$.
Soit $z_{j1}, ...z_{jr}$ sont respectivement des éléments fixés de $\mathcal{Y}_{jr}$.
Un **objet assertion** est une combinaison d'évènements élémentaires qui nous intéressent (ici : $j_1, ..., j_r$) :

$$q = \bigwedge_{\nu=1}^{r} [Y_{j\nu} \mathcal{R}_{j\nu} z_{j\nu}]$$

3. La **fonction d'extension** ("the mapping extension") de l'objet assertion q est une fonction binaire a_q :

$$\begin{array}{rcl} a_q : E & \longrightarrow & \{0,1\} \\ u & \longrightarrow & a_q(u) = \bigwedge_{\nu=1}^{r} [Y_{j\nu}(u) \mathcal{R}_{j\nu} z_{j\nu}] \end{array}$$

4. L'**extension** de l'objet assertion $q = Q = ext(q)$:

$$Q = ext(q) = \{u \in E \mid Y_{j\nu}(u)\mathcal{R}_{j\nu} z_{j\nu} = 1 \quad \forall \nu = 1, ..., r\}$$

Nous pouvons ainsi retrouver parmi les objets réels $u \in E$, ceux qui vérifient les exigences, conditions spécifiées dans la "query" i.e. l'objet assertion q (symbolique virtuel).

Remarque :
Toutes ses définitions sont similaires pour D_j en remplaçant les éléménts z_j par les sous-ensembles D_j.

Exemple :
Reprenons l'exemple du vendeur de voiture (voir 3.5. ; exemple 2).

	$\tilde{Y}_1$	$\tilde{Y}_2$	$\tilde{Y}_3$
1	citroen	22	France
2	citroen	17	France
3	peugeot	20	France
4	peugeot	15	France
5	renault	20	France
6	audi	21	Allemagne
7	mercedes	20	Allemagne
8	mercedes	25	Allemagne
9	bmw	23	Allemagne
10	bmw	24	Allemagne
11	bmw	30	Allemagne
12	fiat	16	Italie
13	fiat	12	Italie
14	fiat	14	Italie
15	alpha roméo	20	Italie

Les variables $\tilde{Y}_1$, $\tilde{Y}_2$, $\tilde{Y}_3$ sont des variables classiques à une seule valeur.
Définissons un objet assertion avec $r = 2$ évènements : $j_1 = 2$ et $j_2 = 3$.
Si $q_1 = [\tilde{Y}_2 \leq 17] \bigwedge [\tilde{Y}_3 \in \{France, Allemagne\}]$, alors les objets satisfaisant q_1 sont : $ext(q_1) = \{2, 4\}$.

4.4 Objets symboliques individuels

4.4.1 Définitions

1. Un **objet symbolique individuel** $q_{(z)}$ est une assertion du type :

$$q_{(z)} = \bigwedge_{j=1}^{p} [Y_j = z_j] = [Y = z]$$

qui est engendrée par un vecteur description $z = (z_1, ..., z_p) \in \chi = \bigotimes_{i=1}^{p} \mathcal{Y}_j$. (i.e. $r = p$ et $\mathcal{R}_j =$ " = " pour $j = 1, ..., p$; une conjonction avec exactement p évènements).

2. **L'extension Q de** $q_{(z)}$ devient :

$$Q = ext(q_{(z)}) = \{u \in E \mid Y_j(u) = z_j \quad \forall j = 1, ..., p\}$$

est l'ensemble de tous les objets de u qui réalisent le même vecteur de données $Y(u) = z$

3. Si un objet assertion q est la conjonction d'exactement p évènements et de type $[Y_j \mathcal{R}_j z_j]$ ou $[Y_j \mathcal{R}_j D_j]$ où chaque variable Y_j, $j = 1, ..., p$ est utilisée une seule fois (comme dans le cas des objets symboliques individuels) alors l'assertion q peut s'exprimer comme :

$$q = \bigwedge_{j=1}^{p} [Y_j \mathcal{R}_j z_j] = [Y \mathcal{R} z]$$

$$q = \bigwedge_{j=1}^{p} [Y_j \mathcal{R}_j D_j] = [Y \mathcal{R} D]$$

où

- la relation $\mathcal{R} = \bigotimes_{j=1}^{p} \mathcal{R}_j$
- $z = (z_1, ..., z_p)$ est un vecteur description
- $D = D_1 \times ..... \times D_p$ est un ensemble description cartésien.

4. Nous pouvons caractériser les cas que nous venons de décrire par une paire $d = (\mathcal{R}, z)$ ou $d = (\mathcal{R}, D)$ qui décrit les exigences pour le vecteur de variables Y.

   Une paire $d = (\mathcal{R}, z)$ ou respectivement $d = (\mathcal{R}, D)$ est appélée **une description**
   où

   - $z \in \chi$ est un vecteur description
   - $D \subseteq \chi$ est un ensemble description
   - $\mathcal{R}$ est une relation "adéquate" appelée aussi un opérateur de comparaison.

5. L'ensemble de toutes les descriptions possibles d d'un problème donné est appelé **l'espace de description** $\mathcal{D}$.

Exemple :

E	$\tilde{Y}_1$	$\tilde{Y}_2$	$\tilde{Y}_3$
citroen	22.000	10	8
peugeot	20.000	15	11
renault	21.000	8	9
audi	25.000	10	12
mercedes	30.000	13	10
bmw	24.000	24	10
fiat	16.000	6	14
alpha romeo	20.000	30	6

Soit un ensemble E de voitures d'occasions :

- $\tilde{Y}_1$ = prix de la voiture (en euros) avec $\mathcal{Y}_1 = I\!R_+$
- $\tilde{Y}_2$= l'âge en mois d'une voiture avec $\mathcal{Y}_2 = I\!R_+$
- $\tilde{Y}_3$= la consommation de carburant (en litre / 100 km) avec $\mathcal{Y}_3 = I\!R_+$.

Nous recherchons une voiture qui vaut au moins 20.000 euros, âgée d'au plus de 24 mois et qui ne consomme pas plus de 10 litres /100 km.

L'ensemble de description cartésien est donc $D = D_1 \times D_2 \times D_3$:

$$[20.000, \infty) \times [0, 24] \times (0, 10] \subseteq \chi = \bigotimes_{j=1}^{3} \mathcal{Y}_j = I\!R_+^3$$

où

$$\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}_3 = " \in "$$

et donc,

$$\mathcal{R} = \bigotimes_{j=1}^{3} \mathcal{R}_j$$

L'objet assertion a la forme suivante :

$$q = [\tilde{Y}_1 \in [20.000, \infty)] \bigwedge [\tilde{Y}_2 \in [0, 24]] \bigwedge [\tilde{Y}_3 (0, 10]]$$

La description $d = (\mathcal{R}, D) = (\bigotimes_{j=1}^{3} " \in ", [20.000, \infty) \times [0, 24] \times (0, 10])$.

L'extension de $q = ext(q) = \{citroen, renault, mercedes, bmw\}$.

Remarque :
Il est possible de réduire des assertions de type $q_1 = [Y \mathcal{R}_1 D]$ en une union d'événements du type $q_2 = [Y \mathcal{R}_2 z]$ et inversément.

4.5 Objets symboliques booléens

4.5.1 Introduction

Dans certains cas pratiques, il est impossible d'exprimer les exigences sous la forme d'objets assertion ou d'objets symboliques individuels. Pour des exigences avec des conjonctions (intersections) et des unions, l'exemple ci-après répertorie les personnes qui auront des problèmes vasculaires plus tard.

$$\begin{aligned} q = [smoking = oui] \quad \bigvee \quad & [[sex = homme] \bigwedge [poids \geq 90] \bigwedge [age \geq 40]] \\ \bigvee \quad & [[sexe = femme] \bigwedge [poids \geq 80] \bigwedge [age \geq 50]] \end{aligned}$$

Pour pouvoir intégrer ces cas dans un modèle de données symboliques, nous allons définir "les Objets symboliques booléens".

4.5.2 Définitions

1. Un **objet symbolique booléen** est une paire $s = (a, d)$
   où
   - $d \in \mathcal{D}$ est une description de l'espace des descriptions $\mathcal{D}$ qui caractérise les propriétés des éléments $u \in E$
   - $a : E \longrightarrow \{0,1\}$ est une fonction binaire qui indique si l'élément $u \in E$ vérifie les propriétés établies :

$$\begin{array}{rcl} a : E & \longrightarrow & \{0,1\} \\ u & \longrightarrow & a(u) = \left\{ \begin{array}{ll} 1 & \text{si } u \text{ vérifie les propriétés} \\ 0 & \text{sinon} \end{array} \right. \end{array}$$

   Les propriétés d'un objet $u \in E$ sont décrites par p variables $Y_1, ..., Y_p$ respectivement à valeurs dans $\mathcal{B}_j$ tel que $\mathcal{B}_j = \mathcal{Y}_j$, $\mathcal{P}(\mathcal{Y}_j)$ ou $\mathcal{M}(\mathcal{Y}_j)$. Le vecteur des données $Y = (Y_1, ..., Y_p)$ prend ses valeurs dans $\mathcal{B} = \bigotimes_{j=1}^{p} \mathcal{B}_j$.

   Si :
   - $\mathcal{B}_j = \mathcal{Y}_j$, j=1,...p et donc que Y est définit dans $\mathcal{B} = \chi = \bigotimes_{j=1}^{p} \mathcal{Y}_j$
   - la description $d = (\mathcal{R}_1, z)$ ou $d = (\mathcal{R}_2, D)$ où
     (a) $z \in \chi$ est un vecteur description
     (b) $D \subset \chi$ est un ensemble description
     (c) $\mathcal{R}_1, \mathcal{R}_2$ sont des relations, des opérateurs de comparaison respectivement sur $\chi \times \chi$ et $\mathcal{P}(\chi) \times \mathcal{P}(\chi)$

   alors l'objet symbolique booléen $s = (Y, d)$ peut s'écrire respectivement sous les formes suivantes :

$$s = (Y, \mathcal{R}_1, z) \quad ; s = [Y \mathcal{R}_1 z]$$

$$s = (Y, \mathcal{R}_2, D) \quad ; s = [Y \mathcal{R}_2 D].$$

2. La **fonction d'extension** ("extension mapping") d'un objet symbolique booléen $s = (Y, d)$ est la fonction binaire :

$$\begin{array}{rcl} a_s : E & \longrightarrow & \{0,1\} \\ u & \longrightarrow & a_s(u) = \left\{ \begin{array}{ll} 1 & \text{si } Y(u) \text{ verifie } d \\ 0 & \text{sinon} \end{array} \right. \end{array}$$

3. Si l'objet symbolique booléen est de la forme $s = (Y, \mathcal{R}_1, z)$
alors

$$a_s(u) = [Y(u)\mathcal{R}_1 z] = \phi(Y(u), z)$$

où $\phi(y, z) = [y\mathcal{R}_1 z] \in \{0, 1\}$ est la fonction binaire définie au point 4.2.2. (similaire pour $s = (Y, \mathcal{R}_2, D)$).

4. L'**extension** d'un objet symbolique booléen $s = (Y, d)$ est définie par :

$$ext(s) = \{u \in E \mid a_s(u) = 1\}$$

Si s=$(Y, \mathcal{R}_1, z)$
alors $ext(s) = \{u \in E \mid [Y(u)\mathcal{R}_1 z] = 1\}$.

4.6 Définitions générales d'objets symboliques - Objets Modaux

4.6.1 Préliminaires

Dans cette section, nous allons exposer la définition générale des objets symboliques. Nous remplacerons la relation $\mathcal{R}$ par la relation "floue" ϕ.
La relation $\mathcal{R}$ ne prend que deux valeurs (1 si la condition est vraie et 0 sinon).
La nouvelle relation ϕ permet de "graduer" par vraie, possible, improbable, faux,... à l'aide de l'intervalle $[0, 1]$.

Remplaçons la fonction binaire $\phi(y, z) = y\mathcal{R}z \in \{0, 1\}$ utilisée pour les objets symboliques booléens par une fonction à valeurs réelles :

$$\begin{aligned} \phi(.,.) : \chi \times \chi &\longrightarrow [0, 1] \\ (y, z) &\longrightarrow \phi(y, z) \end{aligned}$$

telle que $0 \leq \phi(y, z) \leq 1$ indique le degré d'ajustement entre deux vecteurs descriptions y, z de χ.

4.6.2 Définitions générales

1. Un **objet symbolique** est une paire $s = (a, d)$
   où
   - $d \in \mathcal{D}$ est une description de l'espace de description $\mathcal{D}$.
   - la fonction

$$\begin{aligned} a(.) : E &\longrightarrow [0,1] \\ u &\longrightarrow a(u) \end{aligned}$$

   indique le degré avec lequel l'objet u se conforme à la description d et est appelée **la fonction d'extension de** s notée aussi $a_s(.)$

2. **L'extension de niveau** α d'un objet symbolique $s = (a, d)$ dans E est définie par :

$$ext_\alpha(s) = \{u \in E \mid a(u) \geq \alpha\}$$

   où $\alpha \in [0, 1]$ est le niveau ou le seuil.

3. Un objet symbolique $s = (a, d)$ avec une fonction d'extension a non binaire est appelé **un objet modal** .

Exemple :
Soit E = l'ensemble des classes d'espèces de fleurs.
Caractérisons ces espèces par les variables suivantes :

- Y_1 = les couleurs rencontrées dans une espèce.
- Y_2 = les mois de floraisons.

	Y_1	Y_2
{tulipe}	{rouge,jaune}	{juin,juillet}
{rose}	{rose,rouge}	{juin,juillet}
{lilas}	{violet,blanc}	{juillet,août}

Soit l'ensemble de description $D = D_1 \times D_2 = \{jaune, rouge\} \times \{juin, juillet\}$ et $D' = \{rose, rouge\} \times \{juin, juillet\}$, alors $\phi(D, D') = \frac{|D' \bigcap D|}{|D' \bigcup D|} = \frac{2}{6}$ si on considère $\mid D' \bigcap D \mid = 2$ et $\mid D' \bigcup D \mid = 6$.

Chapitre 5

La similarité et la dissimilarité

5.1 Les mesures classiques de similarités et de dissimilarités

5.1.1 Introduction

Les différentes techniques d'Analyse des Données comme la classification sont basées sur les similarités ou dissimilarités qui peuvent exister entre des individus.
L'objectif est de rechercher des classes d'individus $C_1, C_2, ... \subseteq \Omega$ "homogènes" telles que la similarité soit grande entre les paires d'individus d'un même groupe et petite pour des paires d'individus de classes différentes ou telles que la dissimilarité soit petite entre des paires d'individus d'une même classe et grande pour des paires d'individus de groupes différents.

5.1.2 Définitions

Nous caractérisons la **dissimilarité** ou la **similarité** de 2 individus $k, l \in \Omega$ respectivement par une mesure quantitative à valeurs réelles $d(k,l) = d_{kl}$ et $s(k,l) = s_{kl}$.

$$\begin{aligned} dissimilarite : d : \Omega \times \Omega &\longrightarrow \mathbb{R}_+ \\ (k,l) &\longrightarrow d(k,l) = d_{kl} \\ similarite : s : \Omega \times \Omega &\longrightarrow \mathbb{R}_+ \\ (k,l) &\longrightarrow s(k,l) = s_{kl} \end{aligned}$$

Les mesures de dissimilarités et de similarités sont appelées des **mesures de ressemblance.**

5.1.3 Mesures de ressemblance entre objets

Définitions

soit E un ensemble d'éléments (objets) :

1. soit $E = \Omega = \{1, ..., n\}$ un ensemble d'individus
2. soit $E =$ un sous-ensemble de Ω
3. soit $E = \{C_1, ..., C_m\}$ est une collection de classes d'individus $\subseteq \Omega$
4. soit $E =$ l'ensemble des solutions appartenant à $\mathcal{Y}_j$ d'une variable Y_j
5. soit $E = \{x_1, ..., x_n\}$ un ensemble de vecteurs de données qui décrivent les caractéristiques de n individus.

Une **mesure de ressemblance** sur E est une fonction à valeurs réelles $r(a, b)$ qui est définie pour toutes paires (a, b) d'éléménts de E :

$$\begin{aligned} r : E \times E &\longrightarrow \mathbb{R}_+ \\ (k, l) &\longrightarrow r(k, l) \end{aligned}$$

- Si $r(a, b) \equiv d(a, b)$ alors on parlera de **dissimilarité** entre les éléménts a et b.
- Si $r(a, b) \equiv s(a, b)$ alors on parlera de **similarité** entre les éléménts a et b.

Les propriétés pour une mesure de ressemblance r

1. **symétrie** : $\forall a, b \in E \;\; r(a, b) = r(b, a)$
2. $\forall a \in E : r_a^* = r(a, a)$
3. - *dissimilarité* : si $r \equiv d$ alors $r(a, b) = d(a, b) \geq r_a^* \;\forall b \in E$
   - *similarité* : si $r \equiv s$ alors $r(a, b) = s(a, b) \leq r_a^* \;\forall b \in E$

Si $E = \Omega$ est un ensemble d'individus alors les valeurs $r_a^* = r^*$ sont les mêmes $\forall a \in E$.
De plus, $r^* = 1$ dans le cas d'une mesure de similarité et $r^* = 0$ dans le cas d'une mesure de dissimilarité.

Une mesure de dissimilarité d sur E est caractérisée par la condition suivante :

$$0 = d(a,a) \leq d(a,b) = d(b,a) < \infty \quad \forall a, b \in E$$

tandis qu'une mesure de similarité vérifie la condition suivante :

$$1 = s(a,a) \geq s(a,b) = s(b,a) \geq 0 \quad \forall a, b \in E$$

Remarque :
si E est une collection de classes $C_1, ..., C_m \subseteq \Omega$, on peut avoir $d(a,a) > 0$.

Transformation de similarités en dissimilarités et vice versa

Nous pouvons transformer des similarités en dissimilarités et inversement en définissant :

- $d = \phi(s)$ où $\phi(.)$ est une fonction strictement décroissante
- $s = \psi(d)$ où $\psi(.)$ est une fonction strictement décroissante

en imposant respectivement des conditions "frontières" : $\phi(1) = 0$; $\phi(0) = \infty$ et $\psi(0) = 1$; $\psi(\infty) = 0$.
Les transformations ψ les plus utilisées sont :

$$s = max(d) - d \qquad s = \sqrt{max(d) - d} \qquad s = max(d^2) - d^2$$

où $max(d)$ est la valeur maximale observée de d.

Le semi-ordre engendré par une mesure de ressemblance

Une mesure de ressemblance r sur E induit un semi-ordre $\preceq_r$ sur l'ensemble des paires ordonnées $E \times E$:

$$\forall (a,b),(c,d) \in E \times E : (a,b) \preceq_r (c,d)$$

$$\begin{cases} d(a,b) \leq d(c,d) & \text{si } r \equiv d \\ s(a,b) \geq s(c,d) & \text{si } r \equiv s \end{cases}$$

Cela signifie que les éléments de la paire (a,b) se ressemblent plus que les éléments de la paire (c,d).

Propriétés

- *réflexivité* : $\forall a,b \in E : (a,b) \preceq_r (a,b)$
- *transitivité* :

$$\forall (a,b),(c,d),(e,f) \in E \times E$$

$$(a,b) \preceq_r (c,d) \;\; et \;\; (c,d) \preceq_r (e,f) \Rightarrow (a,b) \preceq_r (e,f)$$

- Deux mesures de ressemblance r et r' sont **équivalentes** si et seulement si les ordres correspondants $\preceq_r$ et $\preceq_{r'}$ sur $E \times E$ sont identiques.

5.2 Les fonctions-distance et propriétés spéciales

5.2.1 La matrice de dissimilarités

Définitions

Etant donné qu'il est possible de transformer une mesure de dissimilarité en une mesure de similarité, nous allons restreindre notre discussion aux mesures de disssimilarités uniquement.
Soit d une mesure de dissimilarité sur l'ensemble E qui vérifie la condition suivante :

$$0 = d(a,a) \leq d(a,b) = d(b,a) < \infty \quad \forall a,b \in E \tag{5.1}$$

Si $E = \Omega$ est un ensemble de n individus, nous pouvons définir la **matrice de dissimilarités** $D = (d_{kl})$ k,l=1,...,n
où $d_{kl} = d(k,l) =$ dissimilarité entre deux individus k,l $\in \Omega$.

Propriétés

La matrice de dissimilarité D est :

- *symétrique* : $d_{kl} = d(k,l) = d(l,k) = d_{lk}$
- tout élément de la matrice D est positif : $d_{kl} = d(k,l) \geq 0$
- les éléments diagonaux de D sont égaux à 0.

d est appelée une **"définite dissimilarity"** si :

$$\forall a, b \in E : d(a,b) = 0 \Rightarrow a = b \tag{5.2}$$

Cette propriété permet de conclure que les zéros en dehors de la diagonale sont impossibles.

Inégalité triangulaire :

$$\forall a, b, c \in E : d(a,b) \leq d(a,c) + d(c,b) \tag{5.3}$$

Définitions

1. Une dissimilarité d qui vérifie les conditions (5.1) et (5.3) est appelée une **semi-distance** sur E ou **une pseudo-métrique.**
2. Une dissimilarité d qui vérifie les conditions (5.1),(5.2) et (5.3) est appelée une **distance.**

   $\forall$ d pseudo-métrique, deux éléments $a, b \in E$ tels que $d(a,b) = 0$ ont nécessairement les mêmes dissimilarités avec tous les autres éléments $c \in E$:

   $$d(a,b) = 0 \Rightarrow d(a,c) = d(b,c) \quad \forall c \in E \tag{5.4}$$

3. Une dissimilarité d sur E est une **ultramétrique** si d vérifie (5.1) et l'*inégalité ultramétrique* suivante :

   $$d(a,b) \leq max\{d(a,c), d(c,b)\} \tag{5.5}$$

Remarque :
Un ultramétrique d est une pseudo-métrique ((5.5) $\Rightarrow$ (5.3)).

4. Si les dissimilarités d vérifient (5.1) et l'*inégalité de Buneman* :
$\forall a, b, c, d \in E$

$$d(a,b) + d(c,d) \leq max\{d(a,c) + d(b,d), d(a,d) + d(b,c)\} \qquad (5.6)$$

alors elles sont appelées **des distances-arbre.**

5. Une matrice de dissimilarités $D = (d_{kl})$ $(k, l = 1, ..., n)$ est appelée *Robonsinian* si les dissimilarités d_{kl} augmentent quand k, l s'éloignent de la diagonale $k = l$:

$\forall k \in \{1, ..., n\}$

$$d_{k,k} \leq d_{k,k+1} \leq ... \leq d_{k,n} \quad et \quad d_{k,k} \leq d_{k,k-1} \leq d_{k,k-2} \leq ... \leq d_{k,1}$$

$$d_{k,k} \leq d_{k+1,k} \leq ... \leq d_{n,k} \quad et \quad d_{k,k} \leq d_{k-1,k} \leq d_{k-2,k} \leq ... \leq d_{1,k}$$

5.2.2 Mesures de distance à partir d'une matrice de données classiques

Nous voulons déterminer le degré de similarité ou de dissimilarité entre les paires d'individus k, l d'un ensemble $\Omega = \{1, ..., n\}$ de n individus.
Nous devons construire la matrice de dissimilarité $D = (d_{kl})$ telle que les valeurs d_{kl} pour chaque paire $k, l \in \Omega$ reflètent bien la dissimilarité qu'il existe entre les individus.
La dissimilarité d_{kl} est calculée à partir des vecteurs de données observées $x_k, x_l \in \chi$, les lignes k, l de la matrice des données observées $= \underline{X} = (x_{kj})_{(n \times p)}$.

Les mesures de dissimilarités sont de la forme $d_{kl} = d(x_k, x_l)$
où $d(.,.)$ est une mesure définie sur $\chi = \prod_{j=1}^{p} \mathcal{Y}_j$.
Ces mesures de dissimilarités d (ou de similarités s) peuvent être facilement obtenues à partir de mesures de ressemblance δ_j (dissimilarité) (ou σ_j ; similarité) définies sur $\mathcal{Y}_j$ de chaque variable Y_j $(j = 1, ...p)$ par les formules suivantes :

$$d_{kl} = d(x_k, x_l) = \sum_{j=1}^{p} \delta_j(x_{kj}, x_{lj})$$

$$s_{kl} = s(x_k, x_l) = \sum_{j=1}^{p} \sigma_j(x_{kj}, x_{lj})$$

Nous pouvons également introduire des poids $w_j \geq 0$, $j = 1, ..., p$ dans ces formules :

$$d_{kl} = d(x_k, x_l) = \sum_{j=1}^{p} \omega_j \delta_j(x_{kj}, x_{lj})$$

Les mesures de dissimilarités pour des individus

1. Soit une matrice de données $\underline{X}$ avec p variables quantitatives $x_k, x_l \in \mathbb{R}^p$.

   La **distance euclidienne** sur $\mathbb{R}^p$ est :

   $$d_{kl} = d(x_k, x_l) = \| x_k - x_l \|_2 = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{lj})^2}$$

   qui est une semi-distance sur $\Omega = \{1, ..., n\}$.

2. *Métrique de Minkowski et distance de Manhattan :*
   Soit un nombre réel $q \geq 1$.
   La **distance de Minkowski** ou L_q sur $\mathbb{R}^p$ est une pseudo-distance sur l'ensemble Ω d'individus :

   $$d_{kl} = d_q(x_k, x_l) = \| x_k - x_l \|_q = [\sum_{j=1}^{p} (x_{kj} - x_{lj})^q]^{\frac{1}{q}}$$

   Si q=2, nous retrouvons la distance euclidienne.
   Si q=1, cette distance est connue sous le nom de **distance de Manhattan** :

   $$d_{kl} = d_1(x_k, x_l) = \| x_k - x_l \|_1 = \sum_{j=1}^{p} | x_{kj} - x_{lj} |$$

3. *La distance de Hamming pour des données binaires.*
   Dans le cas de p variables binaires où $x_{kj} \in \{0, 1\}$ $\forall k, j$,
   la métrique de Manhattan est appelée **la distance de Hamming** :

   $$d_{kl} = d(x_k, x_l) = \| x_k - x_l \|_1 = \sum_{j=1}^{p} I_{x_{kj} = x_{lj}} \qquad x_k, x_l \in \{0, 1\}^p$$

Cette formule donne le nombre de composantes identiques $j \in \{1, ..., p\}$ entre les vecteurs binaires x_k, x_l où I est la fonction indicatrice.

4. *La distance euclidienne généralisée et la distance de Mahalanobis.*
Dans le cas de données quantitatives, une généralisation de la distance euclidienne est obtenue en introduisant une matrice B définie positive $(p \times p)$:

$$d_{kl} = \| x_k - x_l \|_{B^{-1}} = \sqrt{(x_k - x_l)' B^{-1} (x_k - x_l)} \quad \forall x_k, x_l \in I\!R^p$$

Cette distance est invariante par rapport à toutes les transformations affines des données $x_k = a + Q x_k$ $(k = 1, ...n)$
où

- $a \in I\!R^p$
- $Q \in I\!R^{p \times p}$ est la matrice de transformation telle que Q'BQ=B.

On utilise cette mesure lorsque nous voulons prendre en compte une dépendance statistique parmi les p variables observées $Y_1, ..., Y_p$.

Si nous choisissons $B = \Sigma$, la matrice de covariance des variables $Y_1, ..., Y_p$ alors la dissimilarité est appelée la **distance de Mahalanobis.**

La matrice Σ est estimée à partir des vecteurs de données $x_1, ...x_n \in I\!R^p$ en utilisant la matrice de dispersion totale T :

$$\Sigma = \frac{1}{n} T = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})(x_k - \overline{x})'$$

où $\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$ est la moyenne de tous les individus.

Dans le cadre de l'analyse de classification, nous pouvons utiliser la matrice de dispersion dans les classes :

$$\Sigma = \frac{1}{n - m} \sum_{i=1}^{m} \sum_{k \in C_i} (x_k - \overline{x}_{C_i})(x_k - \overline{x}_{C_i})'$$

où

- $\mathcal{C} = \{C_1, ..., C_m\}$ est une partition donnée de l'ensemble d'individus dans un certain nombre m de classes $C_1, ..., C_m$
- $\overline{x}_{C_i}$ est la moyenne des vecteurs de données appartenant à la classe C_i.

5.2.3 Mesures de similarité pour les matrices de données-catégories

Mesures de similarité pour des données binaires

Dans le cas de données-catégories, la ressemblance entre deux individus $k, l \in \Omega$ est exprimée par le nombre de composantes de catégories correspondantes ou différentes entre les deux vecteurs de données x_k et x_l.
Dans ce cas-ci, il est souvent d'usage d'utiliser les mesures de similarité à la place des mesures de dissimilarité.

- Soit la matrice de données **binaires** $\underline{X} = (x_{kj})$
  La similarité entre deux lignes x_k et $x_l \in \{0,1\}^p$ peut être calculée à partir des formules suivantes :

$$a = \sum_{j=1}^{p} x_{kj} x_{lj}$$

  a représente le nombre de composantes j identiques et qui valent 1 tel que $x_{kj} = x_{lj} = 1$

$$b = \sum_{j=1}^{p} x_{kj}(1 - x_{lj})$$

  b représente le nombre de composantes j différentes tel que $x_{kj} = 1$ et $x_{lj} = 0$

$$c = \sum_{j=1}^{p} (1 - x_{kj}) x_{lj}$$

  c représente le nombre de composantes j différentes tel que $x_{kj} = 0$ et $x_{lj} = 1$

$$d = \sum_{j=1}^{p} (1 - x_{kj})(1 - x_{lj})$$

d représente le nombre de composantes j identiques et qui valent 0 tel que $x_{kj} = x_{lj} = 0$
où $a + b + c + d = p$.

Les expressions x_{kj} =1 ou 0 interprètent respectivement la présence ou l'absence d'une certaine caractéristique j pour l'individu k.

- Introduisons les similarités les plus souvent utilisées pour des **données binaires :**

1. *Coefficients de similarité et de correspondance :*
   La fréquence relative des composantes identiques pour les deux vecteurs x_k et x_l est souvent utilisée pour mesurer la similarité entre les individus k, l.

   Distinguons deux cas :

   - **les coefficients de correspondance de Sokal-Michener (M-coefficient)**

$$s_{kl} = \frac{a+d}{p} \tag{5.7}$$

   - **le coefficient de similarité de Jaccard (S-coefficient)**

$$s_{kl} = \frac{a}{a+b+c} \tag{5.8}$$

   Nous pouvons remarquer que le M-coefficient est symétrique.
   Si nous inversons toutes les composantes des catégories qui valent 1 et nous les remplaçons par 0 et inversément, la valeur du M-coefficient ne change pas.

2. *La famille de mesures de similarité de Gower-Legendre*
   Gower et Legendre ont introduit des poids $\theta > 0$ dans les formules vues précédemment :

$$S_\theta = \frac{a+d}{a+d+\theta(b+c)} \tag{5.9}$$

$$T_\theta = \frac{a}{a + \theta(b+c)} \tag{5.10}$$

Remarque :
Si $\theta = 1$, nous retrouvons les formules (5.7) et (5.8).
Les mesures de dissimilarités correspondant respectivement aux mesures de similarités (5.9) et (5.10) sont :

$$d = 1 - S_\theta \qquad et \qquad d = 1 - T_\theta$$

Mesures de similarité pour des données-catégories avec plus de deux catégories

Soit une matrice de données-catégories $\underline{X}$
soit p variables Y_j à valeurs respectivement dans $\mathcal{Y}_j = \{1, ..., t_j\}$.

Nous pouvons adapter les formules des données binaires vues précédemment :

1. en transformant chaque vecteur de données x_k (à p composantes) en un vecteur binaire $\tilde{x}_k$ à t composantes où $t = \sum_{j=1}^{p} t_j$ tel que $t_j =\mid \mathcal{Y}_j \mid$.
   t est le nombre total de catégories des p variables.

$$\tilde{x}_k = (\overbrace{0, ..., \underbrace{1}_{x_{k1}}, ..., 0}^{Y_1}; ....; \overbrace{0, ..., \underbrace{1}_{x_{kj}}, ..., 0}^{Y_j}; ....; \overbrace{0, ..., \underbrace{1}_{x_{kp}}, ..., 0}^{Y_p};)$$

   La valeur de la variable Y_j est la catégorie x_{kj}.
   La variable Y_j est un vecteur binaire de longueur t_j avec 1 à la position de la catégorie x_{kj} et 0 ailleurs.

2. en définissant $s(k,l) = s(\tilde{x}_k, \tilde{x}_l)$
   où $s(.,.)$ est une des mesures de similarité pour le cas binaire.

Mesures de ressemblance pour des variables ordinales

Soit Y_j une variable ordinale à valeurs dans $\mathcal{Y}_j = \{1, ..., t_j\}$ contenant t_j catégories ordonnées.
La dissimilarité prend en compte la position des catégories appartenant à $\mathcal{Y}_j$.
Si nous mesurons la dissimilarité entre deux catégories a et b par le nombre

de catégories qu'il y a strictement entre a et b alors nous pouvons coder les t_j categories de $\mathcal{Y}_j$ par les entiers 1,2,3,...,t_j et utiliser n'importe quelle mesure de dissimilarité définie pour des données quantitatives.

Nous avons vu au point 5.2.3 que la mesure de dissimilarité correspondant à la formule de Jaccard (S-coefficient) était $1 - T_1 = \frac{b+c}{a+b+c}$ dans le cas des données binaires.
Généralisons cette formule pour p variables ordinales :

$$d_{kj} = \frac{\sum_{j=1}^{p} x_{kj} + \sum_{j=1}^{p} x_{lj} + 2\sum_{j=1}^{p} min(x_{kj}, x_{lj})}{\sum_{j=1}^{p} x_{kj} + \sum_{j=1}^{p} x_{lj} - \sum_{j=1}^{p} min(x_{kj}, x_{lj})}$$

Formellement, au lieu d'utiliser cette formule, nous pouvons obtenir le même résultat en transformant le vecteur de données x_k en un vecteur binaire x_k^* :

$$x_k^* = (\overbrace{\underbrace{1,1,1,....1,1}_{x_{k1}},0,...,0}^{Y_1};....;\overbrace{\underbrace{1,1,1,...,1,1}_{x_{kj}},0,....,0}^{Y_j};....;\overbrace{\underbrace{1,1,1,...,1,1}_{x_{kp}},.....,0}^{Y_p};)$$

Si l'observation x_{kj} prend les h premières valeurs ordonnées pour la variable Y_j alors les h premières positions du bloc correspondant dans x_k^* sont égales à 1 et les $t_j - h$ positions sont fixées à 0.

Dissimilarités pour des variables mixtes

Si les variables Y_j sont de types différents :

1. Y_j $(j = 1, ..., u)$ les u premières variables sont quantitatives.
2. Y_j $(j = u+1, ..., u+v)$ les v suivantes variables sont de type nominal.
3. Y_j $(j = u+v+1, ..., u+v+w = p)$ les w premières variables sont de type ordinal.

Voici deux méthodes proposées pour résoudre le problème :

1. *Combinaison linéaire des mesures de dissimilarité.*
   Le vecteur de données est divisé en trois parties correspondant aux

trois types de variables :

$$x_k = \begin{pmatrix} x_k^{(Q)} \\ x_k^{(N)} \\ x_k^{(O)} \end{pmatrix}$$

où $x_k^{(Q)}, x_k^{(N)}, x_k^{(O)}$ sont respectivement des vecteurs de données pour les variables quantitatives, nominales et ordinales.
La dissimilarité pour des données mixtes est définie par la formule suivante :

$$d_{kl} = p_1 d^{(Q)}(x_k^{(Q)}, x_l^{(Q)}) + p_2 d^{(N)}(x_k^{(N)}, x_l^{(N)}) + p_3 d^{(O)}(x_k^{(O)}, x_l^{(O)})$$

où

- p_1, p_2, p_3 sont des poids qui reflètent le nombre de variables de chaque type.
- $d^{(Q)}, d^{(N)}, d^{(O)}$ sont respectivement les dissimilarités utilisées pour les variables quantititatives, nominales et ordinales.

2. *Réduction aux mêmes types de données*
Dans cette méthode, les p variables sont transformées dans un même type de variables ; en transformant les p variables soit en variables binaires,soit en p variables quantitatives. Ce genre de manipulation n'est pas toujours possible et contribue souvent à une perte d'information.

Chapitre 6

Les méthodes de classification pour des objets symboliques

6.1 Le problème de classification et des méthodes de classification pour des données classiques

6.1.1 But de la méthode

La tâche principale dans l'analyse des données (classiques ou symboliques) est la détection et la construction de groupes homogènes $C_1, C_2, ...$ d'objets issus d'une population Ω ou E *telle que les objets d'un même groupe montrent une grande similarité* (les objets de groupes diffèrents ont une grande dissimilarité). De tels groupes sont appelés **clusters.**

Cluster analysis est un nom collectif pour les méthodes algorithmiques, statistiques ou mathématiques qui subdivisent l'ensemble Ω ou E en des *clusters* homogènes qui seront rassemblés dans une classification $\mathcal{C} = (C_1, C_2, ...)$.

Les méthodes peuvent être classées selon différents critères tels que :

- le type de données
- le type du critère de classification
- le type de la structure de classification
- le type d'algorithmes.

6.2 Rappels de concepts de base (pour les données classiques)

6.2.1 Le type de données

Soit un ensemble $\Omega = \{1, ..., n\}$ de n objets :

- n vecteurs-données de dimension p : $x_1, ..., x_n$ dans un espace-échantillon χ qui peut être, par exemple, l'espace Euclidien $I\!R^p$ (données quantitatives) ou un ensemble fini de combinaisons de catégories (données qualitatives). Cela correspond à considérer p variables classiques :

$$\tilde{Y}_j : \Omega \to \mathcal{Y}_j \quad j = 1, ..., p$$

qui seront rassemblées dans un vecteur $Y = (\tilde{Y}_1, ..., \tilde{Y}_p)'$.
Alors, les vecteurs $x_k = Y(k) = (\tilde{Y}_1(k), ..., \tilde{Y}_p(k))'$ caractérisent les propriétes des objets $k \in \Omega$.

- Une autre possibilité est de calculer les dissimilarités d_{kl} qui peuvent exister entre les $\begin{pmatrix} n \\ 2 \end{pmatrix}$ paires $k, l \in \Omega$ d'objets.
Les données sont mises dans la matrice de dissimilarité $D = (d_{kl})_{n \times n}$

6.2.2 La structure de classification

Soit les classes $C_1, ..., C_m$ d'une classification $\mathcal{C} = (C_1, ..., C_m)$ de Ω

1. **PARTITION**
Une classification en m classes $\mathcal{C} = (C_1, ...C_m)$ est appelée une **m-partition** de Ω si :
   - $C_i \neq \emptyset \qquad i = 1, ..., m$
   - les classes sont disjointes : $C_i \bigcap C_j = \emptyset \quad i, j = 1, ..., m \quad i \neq j$
   - l'union des classes disjointes est égale à Ω c'est-à-dire que cette classe est exhaustive

$$\bigcup_{i=1}^{m} C_i = \Omega$$

2. **RECOUVREMENT**
Un **recouvrement** de Ω en m classes $C_1, ...C_m$ est défini par :
(a) $C_i \neq \emptyset$ pour $i = 1, ..., m$
(b) $\bigcup_{i=1}^{m} C_i = \Omega$

3. **HIERARCHIES**
Une **hierarchie** $\mathcal{H}$ est une collection finie $\mathcal{H} = (A, B, ...)$ de sous-ensembles $A, B, ... \subset \Omega$ telle que :
- $\Omega \in \mathcal{H}$
- $\forall n$ *singletons* $\{1\}, ...\{n\} \in \mathcal{H}$
- $\forall A, B, ... \in \mathcal{H}$ $A \bigcap B = \emptyset$ *ou* $A \subset B$ *ou* $B \subset A$

Les classes de classification hiérarchique peuvent être représentées par un arbre.

4. **DENDROGRAMME**
Pour chaque classe $A \in \mathcal{H}$, nous pouvons déterminer une fonction monotone h qui mesure l'hétérogénéité d'une classe telle que :

$$h : \mathcal{H} \longrightarrow \mathcal{R}^+$$

vérifie :
- $\forall A, B \in \mathcal{H}, A \subseteq B \Rightarrow h(A) \leq h(B)$ i.e. la classe A est plus homogène ou moins hétérogène que la classe B
- $\forall x_i \in X, h(\{x_i\}) = 0$

Le couple $(\mathcal{H}, h)$ est appelé **un dendrogramme** ou **une hiérarchie indexée.**

Propriétés :
Si n objets sont classés sous la forme d'un dendrogramme
alors,

$$\delta_{kl} = min\{h(A) \mid A \in \mathcal{H}, k \in A, l \in A\} \quad k, l \in \Omega$$

définit le plus petit niveau $h \geq 0$ pour lequel les deux objets k, l appartenant à Ω se rencontreront dans le dendrogramme et caractérise la dissimilarité des objets k, l.

6.3 Partitionnement et critère de clustering

La qualité d'une partition $\mathcal{C} = (C_1, ..., C_m)$ de l'ensemble Ω est souvent mesurée par un critère de classification qui doit être minimisé par rapport à toutes les m-partitions $\mathcal{C}$.

Dans le cas de données quantitatives de dimension p avec les vecteurs $x_1, ..., x_n \in \mathbb{R}^p$, on utilise **le critère de variance intra-classes :**

$$g(\mathcal{C}) := min_{\mathcal{C}} \sum_{i=1}^{m} \sum_{k \in C_i} \| x_k - \overline{x}_{C_i} \|^2$$

où $\overline{x}_{C_i} = \frac{(\sum_{k \in C_i} x_k)}{|C_i|}$ est le centre de la classe C_i

ou bien,

$$g(\mathcal{C}, \mathcal{Z}) := min_{\mathcal{C}, \mathcal{Z}} \sum_{i=1}^{m} \sum_{k \in C_i} \| x_k - z_i \|^2$$

où la minimisation est sur tous les systèmes possibles $\mathcal{Z} := (z_1, ..., z_m)$ avec m centres de classes $z_i \in \mathbb{R}^p$.
Nous pouvons obtenir la minimisation de ce critère en utilisant des algorithmes d'échange (comme par exemple l' algorithme des k-means).

Si les données sont représentées sous la forme d'une matrice $D = (d_{kl})_{n \times n}$ de dissimilarités,
on utilise le critère :

$$W(\mathcal{C}) = min_{\mathcal{C}} \sum_{i=1}^{m} I(C_i)$$

où $I(C_i)$= mesure de l'hétérogénéité de la classe C_i tel que :

$$I(C_i) = \sum_{k \in C_i} \sum_{l \in C_i} d_{kl}$$

ou bien,

$$I(C_i) = \frac{\sum_{k \in C_i} \sum_{l \in C_i} w_k w_l d_{kl}^2}{\sum_{k \in C_i} w_k}$$

où $w_1, ..., w_n$ sont les poids non négatifs pour les éléments de Ω.

6.4 Méthodes de classification

6.4.1 Classement de méthodes de classification

Il existe deux grandes classes de méthodes :

- les méthodes monothétiques
- les méthodes polythétiques.

Les méthodes polythétiques

Les méthodes polythétiques tiennent compte simultanément de l'ensemble des variables décrivant les objets. Ainsi, deux objets pourront appartenir à la même classe sans posséder un seul caractère commun pourvu qu'ils se ressemblent suffisamment du point de vue de l'indice de dissimilarité choisi pour mesurer leur ressemblance. L'information de base manipulée par les méthodes polythétiques est la matrice de proximités et non la matrice de données. Ces méthodes transforment la matrice des proximités en une nouvelle matrice dans laquelle les groupes d'objets sont plus apparents que dans la matrice de proximités initiales. La transformation remplace les dissimilarités initiales par des nouvelles dissimilarités vérifiant des propriétés plus fortes.

Les méthodes monothétiques

Leur objectif est la recherche d'une hiérarchie de partitions , construite à partir de la matrice des données, par une suite de divisions en deux classes ne tenant compte que d'une seule variable à la fois.
A chaque étape, la division peut se faire suivant une variable différente.
Dans ces méthodes, une classe d'objets est définie par la possession en commun d'un attribut. Par exemple, on peut désirer classer l'ensemble de tous les individus qui ont répondu de la même façon à l'une des questions d'une enquête. Le problème se pose alors de sélectionner la question la plus discriminante ou la plus sélective , c'est-à-dire celle qui apporte le plus d'information sur l'ensemble des réponses.

6.4.2 Méthodes hiérarchiques de classification : les méthodes divisives et agglomeratives

L'objectif des méthodes hiérarchiques est la recherche d'une famille de partitions telle que les groupements ou les divisions successifs des objets forment une hiérarchie.

Une classification hiérarchique (H, h) est construite d'une façon récursive :

- soit par une méthode divisive
- soit par une méthode agglomérative.

L'algorithme général pour la méthode agglomérative

On part de n classes-singletons constituées chacune d'un individu : $C_1 = \{a\}, ..., C_n = \{s\}$.
A chaque étape , on regroupe les deux classes les "plus proches" (pour avoir dans une même classe d' objets qui se ressemblent).
On aura donc :

$$\begin{aligned} etape\ 0 &: n\ classes \\ etape\ 1 &: n-1\ classes \\ &\vdots \\ etape\ n-1 &: 1\ classe = \Omega = \{a, ..., s\} \end{aligned}$$

Pour chaque définition différente de la distance entre deux classes, on aura une méthode différente (à chaque étape, les deux classes fusionnées seront différentes).

L'algorithme général pour la méthode divisive

On part d'une classe $\Omega = \{a, .., s\}$.
A chaque étape , on va choisir parmi toutes les divisions possibles d'une classe en deux, celle dont la distance entre les classes obtenues par cette division est maximale jusqu'à ce que une règle convenable d'arrêt empêche des divisions de plus.

On aura donc :

$$\begin{aligned} etape\ 0 &: 1\ classe \equiv \Omega = \{a, ..., s\} \\ etape\ 1 &: 2\ classes \\ &\vdots \\ etape\ n-1 &: n\ classes \end{aligned}$$

- Les méthodes *divisives* **polythétiques** prennent comme entrée une matrice de dissimilarité $D = (d_{kl})_{n \times n}$ qui est construite à partir des p composantes des vecteurs de données simultanément.
- Les méthodes *divisives* **monothétiques** ont comme entrée n vecteurs-données et chaque division est faite en utilisant seulement une seule variable qui est selectionnée (optimalisée).

Pour des données binaires, cela se réduit à une division entre objets qui ont ou qui n'ont pas le niveau, la catégorie d'une variable sélectionnée.
En fait, chaque cluster est caractérisé par une conjonction de propriétés logiques à la fois nécessaires et suffisantes pour être membre du cluster.

Enfin , il existe différentes classes de méthodes hiérarchiques :

1. *Les méthodes hiérarchiques ordinales.*
   Elles n'utilisent pas d'autres informations que le classement des paires d'objets par ordre de proximité.
2. *Les méthodes hiérarchiques non ordinales.*
   Ces méthodes, contrairement aux précédentes, utilisent les valeurs numériques des dissimilarités entre paires d'objets.

Remarques :

1. La hiérarchie peut dépendre de l'ordre d'introduction des données.
2. Le problème lié aux méthodes hiérarchiques est lorsqu'un individu est placé dans un groupe, il y reste jusqu'à la fin. En d'autres termes, il n'est pas possible de corriger une mauvaise partition.
3. Il existe aussi des méthodes hiérarchiques générant une hiérarchie de recouvrements au lieu d'une hiérarchie de partition.

6.5 Méthode de classification divisive pour les données symboliques

Nous proposons dans cette section une méthode divisive de clustering pour un tableau de données symboliques $n \times p : \underline{X}$.
L'algorithme procède de façon monothétique. Cette méthode est définie pour à la fois des données classiques et symboliques.

Nous nous limitons à la présentation de notre méthode au cas de variables symboliques Y_j avec un espace d'observation **ordonné** $\mathcal{Y}_j$ tel que le tableau des données peut contenir des variables classiques, à valeurs multiples, intervalles ou modales.
$\Rightarrow$ Nous ne considérons pas le cas nominal des données.

Une division est définie par l'optimisation d'une généralisation du "critère de variance classique dans les clusters" (i.e. classical winthin-clusters variance criterion) aux cas des données symboliques.
Ce critère est minimisé par rapport à toutes les bipartitions induites par un ensemble de questions binaires relatives aux variables Y_j.

Les clusters ne sont pas systématiquement divisés, mais un des clusters est choisi selon un critère spécifique. Le processus de divisions est stoppé après un nombre d'itérations qui peut être spécifié par l'utilisateur qui est interressé par une classification avec un nombre de clusters petit.
La sortie de la méthode divisive de clustering est une hiérarchie indexée (=dendrogramme) $(\mathcal{H}, h)$ qui est aussi un arbre-décision.

6.5.1 La matrice des données symboliques

Soit un ensemble $\Omega = \{1, ..., n\}$ d'objets caractérisés par p variables symboliques $Y_1, ..., Y_p$:

$$\begin{aligned} Y_j : \Omega &\rightarrow \mathcal{B}_j \\ k &\rightarrow Y_j(k) \end{aligned}$$

où $\mathcal{B}_j$ dépend du type de variable et de $\mathcal{Y}_j$.

- Si $\mathcal{Y}_j = I\!R$ alors on considère deux types de variables Y_j :
  - variable quantitative Y_j avec $Y_j(k) \in I\!R$, donc $\mathcal{B}_j = I\!R$.
  - variable (symbolique) intervalle Y_j avec $Y_j(k) = [\alpha, \beta] \subset I\!R$ donc $\mathcal{B}_j$ est un ensemble des intervalles fermés et bornés $\mathcal{I}$ dans $I\!R$.
- Si $\mathcal{Y}_j = \{a, b, c, ..., h\}$ est un ensemble fini de catégories totallement ordonnées alors on considére trois types de variables :
  - variable ordinale classique Y_j où $Y_j(k) \in \mathcal{Y}_j$ et $\mathcal{B}_j = \mathcal{Y}_j$
  - variable multivariée Y_j où $Y_j(k) \subset \mathcal{Y}_j$ et $\mathcal{B}_j = \mathcal{P}(\mathcal{Y}_j)$
  - variable modale Y_j où $Y_j(k) = \pi_k$ est une distribution de probabilité de fréquence sur $\mathcal{Y}_j$ et $\mathcal{B}_j$ est l'ensemble $\mathcal{M}(\mathcal{Y}_j)$ de toutes les distributions de probabilité sur $\mathcal{Y}_j$.

Finallement, un objet $k \in \Omega$ est décrit par le vecteur des variables symboliques $Y = (Y_1, ..., Y_p)'$ avec des observations :

$$Y(k) = (Y_1(k), ..., Y_p(k))' = \xi_k = (\xi_{k1}, ..., \xi_{kp})'$$

Ces "valeurs" sont rangées dans le tableau des données symboliques :

$$\underline{X} = \begin{pmatrix} \xi_1' \\ \cdot \\ \cdot \\ \cdot \\ \xi_n' \end{pmatrix}$$

La méthode de clustering ne permet pas de mélanger n'importe quel type de variables. Nous supposons que les $\mathcal{Y}_j$ sont tous égaux à $I\!R$ (des variables quantitatives) ou à un ensemble de catégories.
On ne peut pas mélanger les variables continues (quantitatives classiques ou intervalles) avec les variables discrètes (nominales classiques, multivaluées ou probabilistes). De plus, les résultats n'ont un sens que si les modalités des variables discrètes sont ordonnées.
Par exemple, les modalités non ordonnées bleu, blanc et vert donneront des résultats faux tandis que les modalités ordonnées petit, grand et moyen, donneront des résultats convenables.
De plus, les modalités doivent être bien déclarées dans l'ordre croissant.

6.5.2 Deux mesures de distances

Nous proposons deux mesures de distance sur l'ensemble $\Omega = \{1, ..., n\}$ pour déterminer la matrice de distance $D = (d_{kl})_{n \times n}$ issue de la matrice symbolique $\underline{X}$.

Une mesure de distance d'une matrice de données quantitatives (symboliques)

Nous combinons p indices de dissimilariés $\delta_1, ..., \delta_p$ définis sur $\mathcal{B}_j$ en une mesure de dissimilarité globale sur Ω en utilisant une formule de composition.

- Soit δ_j une fonction-distance définie sur $\mathcal{B}_j$

$$\begin{array}{rcl} \delta_j : \mathcal{B}_j \times \mathcal{B}_j & \rightarrow & I\!R^+ \\ (\xi_k^j, \xi_l^j) & \rightarrow & \delta_j(\xi_k^j, \xi_l^j) \end{array}$$

  Si $\xi_k^j = [\alpha_k^j, \beta_k^j]$ et $\xi_l^j = [\alpha_l^j, \beta_l^j]$ sont des intervalles alors nous utilisons la distance de Hausdorff :

$$\delta_j(\xi_k^j, \xi_l^j) = max\{\mid \alpha_k^j - \alpha_l^j \mid, \mid \beta_k^j - \beta_l^j \mid\}$$

  dans le cas de variables à une valeur, la formule est réduite à la valeur absolue entre deux valeurs-singletons.

- Les distances spécifiques δ_j pour chaque variable sont combinées en une fonction distance d définie sur Ω par la formule de composition :

$$\begin{array}{rcl} d : \Omega \times \Omega & \rightarrow & I\!R_+ \\ (k, l) & \rightarrow & d(k, l) = (\sum_{j=1}^{p} [\delta_j(\xi_k^j, \xi_l^j)]^2)^{\frac{1}{2}} \end{array}$$

  donc, pour les variables intervalles, on obtient une distance non normalisée :

$$\begin{array}{rcl} d : \Omega \times \Omega & \rightarrow & I\!R_+ \\ (k, l) & \rightarrow & d(k, l) = (\sum_{j=1}^{p} [max\{\mid \alpha_k^j - \alpha_l^j \mid, \mid \beta_k^j - \beta_l^j \mid\}]^2)^{\frac{1}{2}} \end{array}$$

  Si les intervalles se ramènent à un point, on obtient la distance euclidienne sur $I\!R^p$

Nous combinons maintenant les fonctions δ_j avec une autre formule de composition et on obtient une fonction-distance normalisée :

$$\begin{aligned} d : \Omega \times \Omega &\rightarrow \mathbb{R}_+ \\ (k,l) &\rightarrow d(k,l) = (\sum_{j=1}^{p} [\frac{\delta_j(\xi_k^j, \xi_l^j)}{m_j}]^2)^{\frac{1}{2}} \end{aligned}$$

Deux normalisations sont possibles :

1. soit m_j est l'extension de la déviation standard au cas d'une variable symbolique définie par :

$$m_j^2 = \frac{1}{2n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} (\delta_j(\xi_k^j, \xi_l^j))^2$$

2. soit m_j est la longueur du domaine $\mathcal{Y}_j$

Mesure de distance d'une matrice de données (symboliques) catégories

La matrice des données symboliques est codée comme une matrice de données de fréquences $\underline{X} = (f_{kj})_{n \times m}$.
Si Y_j est une variable à valeurs multiples, l'ensemble des catégories $Y_j(k)$ est codé comme une distribution de fréquence où $m = \mid Y_1 \mid + \ldots + \mid Y_p \mid$ représente le nombre de catégories.

Pour comparer deux objets k et l de Ω, nous utilisons comme la distance :

$$d^2(k,l) = \sum_{j=1}^{m} \frac{p_{..}}{p_{.j}} (\frac{p_{kj}}{p_{k.}} - \frac{p_{lj}}{p_{l.}})^2$$

où

$$\begin{aligned} p_{kj} &= \frac{f_{kj}}{np} \\ p_{k.} &= \sum_{j=1}^{m} p_{kj} \\ p_{.j} &= \sum_{k=1}^{n} p_{kj} \\ p_{..} &= \sum_{k=1}^{n} \sum_{j=1}^{m} p_{kj} = 1 \end{aligned}$$

6.6 L'extension du critère de variance intra-classes (within-class)

soit $D = (d_{kl})_{n\times n}$ la matrice de distance définie sur $\Omega = \{1, ...n\}$. Chaque objet a un poids $w_k \geq 0$ $(k = 1, ..., n)$. Le critère utilisé pour évaluer la qualité d'une partition est une extension du critère $g(\mathcal{C})$ vu au point 6.3..

Définition :

- **Une extension** de la formule suivante :

$$\tilde{I}(C_i) = \sum_{k\in C_i} \| x_k - \overline{x}_{C_i} \|^2$$

d'un cluster C_i des données quantitatives classiques au cas de la matrice-distance $D = (d_{kl})_{n\times n}$ est donnée par :

$$I(C_i) = \frac{1}{2\mu_i} \sum_{k\in C_i} \sum_{l\in C_i} w_k w_l d_{kl}^2 \tag{6.1}$$

avec les poids $\mu_i = \sum_{k\in C_i} w_k$.

Maintenant, nous utilisons $w_k = \frac{1}{n}$ et donc

$$I(C_i) = \frac{1}{n \times n_i} \sum_{k=1}^{n_i-1} \sum_{l>k}^{n_i} d_{kl}^2 \tag{6.2}$$

où $n_i = \mid C_i \mid$ est le nombre d'objets de C_i.

- **La généralisation du critère de variance** $g(\mathcal{C})$ d'une m-partition $\mathcal{C} = (C_1, ..., C_m)$ au cas d'une matrice-distance $D = (d_{kl})_{n\times n}$ est donnée par :

$$W(\mathcal{C}) = \sum_{i=1}^{m} I(C_i)$$

où $I(C_i)$ est la mesure de l'hétérogénéité (formules (6.1.) ou (6.2.)).

6.7 Le bipartionnement d'un groupe

Soit C_i un ensemble de n_i objets.
Le but est de trouver la bipartition $C_i = (C_i^1, C_i^2)$ avec la plus petite variance intra-classe. Une des méthodes était de choisir la bipartition (C_i^1, C_i^2) optimale parmi les $2^{n_i-1} - 1$ bipartitions possibles. Mais il est clair que le coût des calculs est trop élevé quand n_i est grand ($\Longrightarrow$ méthode impossible à appliquer).
Nous réduisons la complexité en choisissant la meilleure partition parmi toutes les partitions induites dans l'ensemble des questions binaires possibles.

6.7.1 Question binaire et données symboliques

Dans le cadre de la méthode divisive de clustering,

- *avec des données quantitatives (valeur simple)* :
  Un cluster C est divisé selon une question binaire de la forme :

  $$Is\ Y_j \leq c?$$

  où $c \in \mathcal{Y}_j$ est appelé **valeur de coupure**

- *avec des données symboliques*
  Un objet $k \in C$ répond "oui" ou "non" à la question binaire selon une fonction binaire :

  $$q_c : \Omega \rightarrow \{true, false\}$$

  et la bipartition (C_1, C_2) de C induite par la question binaire "$Is\ Y_j \leq c$? est comme ceci :

  $$\begin{aligned} C_1 &= \{k \in C \mid q_c(k) = true\} \\ C_2 &= \{k \in C \mid q_c(k) = false\} \end{aligned}$$

Plus spécifiquement, dans le cas des données symboliques, nous considèrons les dichotomies suivantes :

1. **Pour une variable intervalle** $Y_j(k) = [\alpha, \beta]$
   q_c est définie comme :

$$q_c(k) = \begin{cases} true & \text{si } m_k \leq c \\ false & \text{si } m_k > c \end{cases}$$

   où $m_k = \frac{\alpha+\beta}{2}$ est le milieu de l'intervalle $[\alpha, \beta]$.

2. **Pour une variable modale ou à valeurs dans un ensemble** $Y_j(k) = \pi_k$ est sans restriction une probabilité discrète ou une distribution de fréquence. En effet, une variable à valeur dans un ensemble peut être codée comme une variable modale en définissant π_k comme une distribution de probabilité uniforme sur l'ensemble des catégories observées. La fonction q_c est définie comme :

$$q_c(k) = \begin{cases} true & \text{si } \sum_{x \leq c} \pi_k(x) \geq \frac{1}{2} \\ false & \text{si } \sum_{x \leq c} \pi_k(x) < \frac{1}{2} \end{cases}$$

6.7.2 Choix de la meilleure bipartition

Le critère $g(\mathcal{C})$ est minimisé parmi toutes les bipartitions induites par l'ensemble de toutes les questions binaires.
Soit z_j le nombre de bipartitions induites par la variable Y_j.

- Si Y_j est une **variable intervalle**, il y aura au plus $z_j = n_i - 1$ différentes bipartitions (C_i^1, C_i^2) induites par cette variable.
  En effet, chaque fois que la valeur de coupure c peut être entre deux m_k consécutifs, la division induite est la même.

  Nous décidons d'utiliser les $n_i - 1$ valeurs de coupures c, choisies comme le milieu des m_k consécutifs. En effet, si les n_i valeurs m_k sont différentes, il y a $n_i - 1$ valeurs de coupure sur Y_j.

- Si Y_j est une **variable décrite par une probabilité ou une distribution de fréquence $\mathcal{Y}_j$ (fini ou ordonnée)** et que $m_j = \mid \mathcal{Y}_j \mid$, alors il y a $m_j - 1$ différentes questions binaires et au plus $z_j = m_j - 1$ différentes bipartitions (C_i^1, C_i^2) induites par cette variable.

$\rightarrow$ Si il y a p variables, nous choisissons parmi les $z_1 + ... + z_p$ bipartitions (C_i^1, C_i^2) la partition avec la plus petite inertie.

6.7.3 Choix du cluster à diviser

Soit $P_m = (C_1, ..., C_m)$ une m-partition de Ω.
A chaque étape, une nouvelle $(m+1) - partition$ est obtenue en divisant un des clusters $C_i \in P_m$ en deux nouveaux clusters C_i^1 et C_i^2
Le problème est de choisir le cluster $C_i \in P_m$ tel que la nouvelle partition

$$P_{m+1} = P_m \bigcup \{C_i^1, C_i^2\} - \{C_i\}$$

a une variance intra-classes minimum.
Nous savons que $W(P_{m+1}) = W(P_m) - I(C_i) + I(C_i^1) + I(C_i^2)$
Dans cette expression, minimiser $W(P_{m+1})$ est équivalent à choisir le cluster $C_i \in P_m$ tel que la différence entre l'inertie de C_i et de l'inertie des bipartitions (C_i^1, C_i^2) est maximum.
Donc, le critère de sélection du cluster à diviser est donné par :

$$\Delta(C_i) = I(C_i) - I(C_i^1) - I(C_i^2) \tag{6.3}$$

6.7.4 La règle d'arrêt et la sortie

Le processus de divisions est stoppé après L itérations et L est donné comme entrée par l'utilisateur.
Le problème d'arrêt du processus de division avant d'obtenir la partition la plus fine (avec n singletons ; $L = n$) est d'assurer que les partitions intéressantes avec une petite variance ont déjà été détectées après $L < n$ itérations. Cette propriété est vérifiée parce que les groupes ne sont pas systématiquement divisés, mais un des clusters est choisi selon le critère (6.3.) qui s'assure que la partition induite par cette division a une variance minimum.
La sortie de la méthode divisive est une hiérarchie $\mathcal{H}$ dont les singletons sont

les $L+1$ clusters de la partition obtenue dans la dernière itération de l'algorithme. Chaque cluster $C_i \in \mathcal{H}$ est indexé par $\Delta(C_i)$ dans le dendrogramme car

$$C_i \subset C_{i'} \Rightarrow \Delta(C_i) \leq \Delta(C_{i'})$$

Il n' y aura pas d'inversion dans le dendrogramme de la hiérarchie.
Cette hiérarchie est un arbre de décision. Les L clusters sont les feuilles et les noeuds sont les questions binaires selectionnées par l'algorithme. Chaque cluster est caractérisé par une règle définie selon les questions binaires conduisant de la racine aux feuilles correspondantes.

Chapitre 7

Vue d'ensemble de l'analyse factorielle

7.1 Introduction

L'analyse factorielle simple permet l'exploration globale d'un tableau de données statistiques décrivant des individus grâce à diverses variables de même nature (toutes quantitatives ou qualitatives).
L'analyse en composantes factorielles tente de représenter les variables $Y_1, ...Y_p$ comme une combinaison linaire de quelques variables aléatoires appeles **facteurs** $f_1, ..., f_m$ où $m < p$.
Il est à noter que ces facteurs ne sont ni observables, ni mesurables.

7.2 Définitions et hypothèses

Considérons donc p variables observées $Y_1, ..., Y_p$ qui ont pour vecteur moyenne μ et comme matrice de covariance Σ.
Nous devons exprimer chaque variable par une combinaison linéaire des facteurs communs $f_1, ..., f_m$ $(m < p)$ avec une certaine erreur qui explicite la partie spécifique de chaque variable.

Nous obtenons donc le système d'équations suivant :

$$\begin{aligned} Y_1' &= Y_1 - \mu_1 = l_{11}f_1 + l_{12}f_2 + ... + l_{1m}f_m + \varepsilon_1 \\ &\vdots \\ Y_p' &= Y_p - \mu_p = l_{p1}f_1 + l_{p2}f_2 + ... + l_{pm}f_m + \varepsilon_p \end{aligned}$$

où

- Y_i' est la variable i réduite qui est exprimée par des facteurs f_j $(j = 1, ..., m)$ aléatoires, non observés.
- l_{ij} sont les poids qui montrent la proportion de dépendance des variables Y_i par rapport aux facteurs f_j.
- ε_i représente l'erreur expliquant la partie spécifique de chaque variable.

Nous supposerons que les facteurs vérifient les conditions suivantes :

- $E(f_j) = 0 \ \forall j = 1, ..., m$
- $Var[f_j] = 1 \ \forall j = 1, ..., m$ i.e. les facteurs communs sont centrés réduits comme les variables
- $Cov(f_i, f_j) = 0$ pour $i \neq j$.

Nous supposerons également :

- $E(\varepsilon_i) = 0 \ \forall i = 1, ..., p$
- $Var(\varepsilon_i) = \psi_i$ car les facteurs sont spécifiques à chaque variable et ont donc des variances différentes. Nous appellerons ψ_i la variance spécifique.
- $Cov(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j$

Ajoutons une dernière hypothèse qui montre que les ε_i décrivent la partie non commune des variables :

$$Cov(\varepsilon_i, f_j) = 0 \quad \forall i, j$$

Nous pouvons également facilement écrire la variance des variables avec les hypothèses précédentes :

$$Var(Y_i) = l_{i1}^2 + l_{i2}^2 + l_{i3}^2 + ... + l_{im}^2 + \psi_i = 1$$

où

- $\psi_i = var(\varepsilon_i)$ représente la variance spécifique
- $\sum_{j=1}^{m} l_{ij}^2 = h_i^2$ représente la variance commune.

Cette variance globale est égale à l'unité par le choix de variables centrées réduites.
La variance commune mesure la part de variabilité totale qui est expliquée par l'ensemble des facteurs communs f_j et la variance spécifique mesure la part de variabilité totale qui est expliquée par le facteur spécifique correspondant.

Notations matricielles

Nous pouvons, pour faciliter l'écriture, utiliser les notations matricielles, nous obtenons ainsi l'expression suivante :

$$Y' = Y - \mu = Lf + \varepsilon$$

où

- $Y = (Y_1, ..., Y_p)'$ est le vecteur de variables
- $Y' = (Y_1', ..., Y_p')'$ est le vecteur des variables réduites
- $\mu = (\mu_1, ..., \mu_p)'$ est le vecteur moyenne
- L est la matrice ($p \times m$) des poids des facteurs communs :

$$L = \begin{pmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \vdots & \vdots \\ l_{p1} & \cdots & l_{pm} \end{pmatrix}$$

- $f = (f_1, f_2, ..., f_m)'$ est le vecteur des m facteurs communs
- $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)$ est le vecteur des p facteurs spécifiques

Les hypothèses deviennent :

- $E(f) = 0$
- $E(\varepsilon) = 0$
- $Cov(f, \varepsilon) = 0$
- $Cov(\varepsilon) = \psi = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$
- $Cov(f) = I$, la matrice identité

La matrice de covariance Σ peut s'exprimer en terme de ψ et L :

$$\Sigma = LL' + \psi' \tag{7.1}$$

7.3 Estimation des poids et des facteurs

Toute l'analyse en composantes factorielles réside dans :

- le calcul des poids l_{ij}
- l'estimation des facteurs f_j.

7.3.1 Estimation et calcul des poids

Le but de l'analyse en composantes factorielles est de calculer les poids de L qui permettent d'expliquer au mieux, avec le plus petit nombre de facteurs communs, les fluctuations des variables initiales.

Il existe différentes méthodes :

1. Méthode en composantes principales
2. Méthode des facteurs principaux
3. Méthode du maximum de vraisemblance

Nous ne développerons ici que la méthode en composantes principales.

7.3.2 Méthode en composantes principales

De l'échantillon de variables $Y_1, Y_2, Y_3, ..., Y_p$, nous retirons une matrice de covariance que nous noterons S. Il faut que nous trouvions un estimateur $\hat{L}$ pour approximer la structure que nous avons rencontrée (7.1.),c'est-à-dire avoir l'expression :

$$S \cong \hat{L}\hat{L}' + \hat{\psi}$$

Dans cette méthode, nous négligeons la partie $\hat{\psi}$ et nous gardons la factorisation $S = \hat{L}\hat{L}'$.
Nous opérons une décomposition spectrale qui nous donne :

$$SC = CD \quad ou \quad S = CDC'$$

où

- C est une matrice orthogonale construite avec les vecteurs propres de S, ses vecteurs propres étant normalisés : $C_i'C_i = 1$.
- D est la matrice des valeurs propres de $S = diag(\lambda_1, \lambda_2, ..., \lambda_p)$.

Pour arriver à notre but, il faudrait pouvoir identifier $S = CDC'$ à $\hat{L}\hat{L}'$. Or, la matrice S est définie positive et donc ceci nous permet d'écrire D comme le produit suivant :

$$D = D^{\frac{1}{2}}D^{\frac{1}{2}}$$

où $D^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_p})$

et donc,

$$\begin{aligned} S &= CDC' \\ &= CD^{\frac{1}{2}}D^{\frac{1}{2}}C' \\ &= (CD^{\frac{1}{2}})(CD^{\frac{1}{2}})' \end{aligned}$$

car $D^{\frac{1}{2}}$ est une matrice diagonale.

Nous pourrons choisir comme solution $\hat{L} = CD^{\frac{1}{2}}$
Cependant, cette matrice $CD^{\frac{1}{2}}$ est de dimension $p \times p$ alors que la matrice

$\hat{L}$ est de dimension $p \times m$ avec $m < p$.
Nous devons donc réduire les dimensions de $CD^{\frac{1}{2}}$.

Définissons les matrices C_1 et D_1 :

- la matrice D_1 contient les m plus grandes valeurs propres de S telles que $\lambda_1 > \lambda_2 > ... > \lambda_m$.
- la matrice C_1 contient les m vecteurs propres correspondants.

Nous pouvons donc affecter : $\hat{L} = C_1 D_1^{\frac{1}{2}}$
De plus, la somme des carrés des lignes de $\hat{L}$ est égale à la variance commune :

$$\sum_{j=1}^{m} \hat{l}_{ij}^2 = h_i^2$$

Remarque :
Les poids des facteurs ne sont pas uniques car il est possible de multiplier les poids par une matrice orthogonale sans modifier la matrice de covariance. La multiplication d'un vecteur par une matrice orthogonale est équivalente à une rotation des axes.

7.3.3 Estimation des facteurs

L'estimation des facteurs a pour but de déterminer les valeurs des variables fondamentales communes f_k relatives aux différents individus considérés.
Ces valeurs constituent une matrice F de dimension $m \times n$ et peut être obtenue par régression multiple à partir des valeurs observées réduites centrées :

$$F = BX'$$

où

- B = la matrice de régression $(m \times p)$
- X' = matrice des valeurs observées centrées réduites $(p \times n)$

7.4 Les rotations

Très souvent, les facteurs communs f_k ne peuvent être interprétés facilement. Alors, on essaie de remplacer ces facteurs par d'autres solutions équivalentes

permettant une meilleure interprétation.
Nous avons vu plus haut que les poids n'étaient uniques qu'à travers la multiplication d'une matrice orthogonale.
Ainsi, la matrice des poids estimée $\hat{L}$ peut être transformée par une rotation.
Le but d'une rotation est de placer les axes "les plus près possible " des groupes de points afin de rendre l'interprétation plus objective.

Il existe deux grands types de rotations :

- les rotations **orthogonales**
- les rotations **obliques.**

7.4.1 Les rotations orthogonales

Les axes de rotations sont orthogonaux. Nous multiplions la matrice des poids par une matrice orthogonale. Les rotations orthogonales préservent les angles, les distances ainsi que les variances communes.
La méthode la plus connue est la méthode Varimax.
Cette méthode de rotation est habituellement utilisée pour donner à la matrice des poids L une structure plus simple.
Le critère de cette méthode est de maximiser la quantité suivante :

$$\sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{p} \frac{\hat{l}_{ij}^4}{\hat{h}_i^4}}{p} \right) - \left(\frac{\sum_{i=1}^{p} \frac{\hat{l}_{ij}^2}{\hat{h}_i^2}}{p} \right)^2$$

qui représente la somme des variances de chaque $\frac{\hat{l}_{ij}^2}{\hat{h}_i^2} \ \forall j$

7.4.2 les rotations obliques

Dans ce cas-ci, les axes de rotations ne sont pas orthogonaux.
De plus, les rotations obliques n'utilisent plus une matrice de transformation orthogonale mais elles utilisent une matrice de transformation générale non-singulière.
Exemples de rotations obliques : quartimax, orthomax,....
Nous ne developperons pas ces méthodes ici.

Deuxième partie

Recherche et résultats

Chapitre 8

Les fichiers

8.1 Les fichiers de données

8.1.1 Introduction

Nous disposons de six fichiers de données :

- *communes.xls*
- *popact.xls*
- *agelog.xls*
- *tailmen.xls*
- *popage98.xls*
- *revenu.xls*

Tous ces fichiers nous ont été transmis par le Service des Etudes et de la Statistique de la Région Wallonne.

Le premier fichier *communes.xls* est un fichier ne contenant que des variables quantitatives tandis que les autres fichiers ne contiennent que des variables qualitatives. Etant donné que la méthode divisive de classification ne peut traiter les variables quantitatives et qualitatives en même temps, les recherches que nous avons réalisées se décomposent en deux parties.
La première concerne le fichier de données *communes.xls*.
La deuxième partie comprend les recherches effectuées sur les 5 autres fichiers.

8.1.2 Le fichier *communes.xls*

Ce fichier comprend 262 communes. Chaque commune est caractérisée par 12 variables quantitatives.

- INS est le code des communes wallonnes.
- $VARPOP$ représente la variance de la population de chaque commune.
- $ETAB500$ assure la présence d'un établissement employant plus de 500 personnes.
- $ENSSUP$ assure la présence d'un établissement d'enseignement supérieur.
- $POP75$ donne le pourcentage de la population de plus de 75 ans.
- $MINIMEX$ représente le pourcentage de minimexés.
- $TAIL.MEN$ donne la taille moyenne des ménages.
- $BATI.LOG$ représente le pourcentage de bâtiments sur la zone à bâtir de la commune.
- $CONFORT$ donne l'indice de confort de la population d'une commune.
- $OFFRE.EMPL$ représente les offres d'emplois satisfaites dans la commune.
- $POP.ACT$ donne le pourcentage de la population active de la commune.
- $TYPO.EVOLPOP$ représente le type d'évolution de la population (quatre catégories différentes).

Voici un échantillon de 3 communes du fichier *communes.xls* pour permettre une meilleure visualisation de la forme du fichier de données :

	INS	VARPOP	ETAB500	ENSSUP	POP75
Beauvechain	25005	4.52	1	0	5.95
Braine-L'alleud	25014	18.46	1	0	5.08
Mons	53053	19.55	1	2	6.35

	MINIMEX	TAILMEN	BATI.LOG
Beauvechain	4.45	2.81	86.5
Braine-L'alleud	3.74	2.74	79.6
Mons	22.73	2.33	62.4

	CONFORT	OFFRE.EMPL	POP.ACT	TYPO.EVOLPOP
Beauvechain	69	27.76	42.62	1
Braine-L'alleud	76	26.11	45.81	3
Mons	56	34.33	36.55	2

Dans notre analyse, nous n'avons pas considéré les variables : INS, ETAB500, et ENSSUP. En effet, la variable INS représente le code de la commune et les variables ETAB500 et ENSSUP sont des variables binaires. En conclusion, nous ne considérons dans notre étude que les huit variables qui sont des variables quantitatives continues.

8.1.3 Les fichiers *popact.xls, agelo.xls, tailmen.xls, popage98.xls, revenu.xls*

Description des différents fichiers

1. Le fichier *popact.xls* contient une variable décrivant la structure des activités de la population. Cette variable regroupe 9 catégories.

   - S_1 : Le nombre d'effectifs de la population active occupés.
   - S_2 : Le nombre d'effectifs de la population active non occupés et demandeurs d'emploi.
   - S_3 : Le nombre d'effectifs de la population active non occupés et les miliciens.
   - S_5 : Le nombre d'effectifs de la population non active de moins de 18 ans.
   - S_6 : Le nombre d'effectifs de la population non active de plus de 18 ans aux études.
   - S_7 : Le nombre d'effectifs de la population non active s'occupant du ménage.
   - S_8 : Le nombre d'effectifs de la population non active ayant cessé le travail.
   - S_9 : Le nombre d'effectifs de la population non active pour raison non définie.
   - S_11 : Le nombre d'effectifs de la population avec un statut indéterminé.

Voici un extrait du fichier *popact.xls* :

	S_1	S_2	S_3	S_5	S_6	S_7	S_8	S_9	S_{11}
Beauvechain	2233	209	16	1362	295	693	814	45	100
Braine-L'alleud	13523	1216	129	7709	1846	3283	4185	244	323
Mons	26543	6718	264	20360	4189	14362	12718	880	5692

2. Le fichier *agelo.xls* contient une variable décrivant la répartition des constructions de logements suivant des périodes de temps. Cette variable regroupe 8 catégories.
   - C_2 : Le nombre de logements construits avant 1919.
   - C_3 : Le nombre de logements construits entre 1919 et 1945.
   - C_4 : Le nombre de logements construits entre 1945 et 1961.
   - C_5 : Le nombre de logements construits entre 1962 et 1970.
   - C_6 : Le nombre de logements construits entre 1971 et 1980.
   - C_7 : Le nombre de logements construits entre 1981 et 1985.
   - C_8 : Le nombre de logements contruits en 1986 et après.
   - C_9 : Le nombre de logements construits dont l'année de construction n'a pas été spécifiée.

   Voici un extrait du fichier agelo.xls :

	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Beauvechain	740	211	251	127	332	72	98	154
Braine-L'alleud	1630	1130	1447	1523	3517	567	739	1114
Mons	8977	4866	4943	3367	3567	866	679	6540

3. Le fichier *tailmen.xls* contient une variable décrivant la taille des ménages. Cette variable regroupe 8 catégories.
   - T_1 : Le nombre de ménages composés d'un homme seul.
   - T_2 : Le nombre de ménages composés d'une femme seule.
   - T_3 : Le nombre de ménages composés de 2 personnes.
   - T_4 : Le nombre de ménages composés de 3 personnes.
   - T_5 : Le nombre de ménages composés de 4 personnes.

- T_6 : Le nombre de ménages composés de 5 personnes.
- T_7 : Le nombre de ménages composés de 6 personnes.
- T_8 : Le nombre de ménages composés de 7 personnes.
- T_9 : Le nombre de ménages composés de 8 personnes et plus.

Voici un échantillon de trois communes du fichier *tailmen.xls* :

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
Beauvechain	167	228	567	452	387	172	46	19
Braine-l'alleud	977	1562	3329	2575	2274	877	230	39
Mons	5712	8098	11246	7243	4861	1711	499	134

4. Le fichier *popage98.xls* reprend la répartition de la population suivant l'âge, le sexe et la nationalité des personnes se trouvant sur le territoire des communes wallonnes.
   Les catégories POP_1,...,POP_{20} donnent le nombre de personnes se trouvant respectivement dans les différentes classes d'âge : 0-4 ans, 5-9 ans, 10-14 ans, ..., 85-89 ans, 90-94 ans, 95 ans et plus.
   La variable $SEXE$ désigne le sexe des personnes : 1 pour les hommes et 2 pour les femmes.
   La variable $BELETR$ designe la nationalité des personnes : 1 pour les belges et 2 pour les étrangers.

   Chaque commune donne une répartition de la population suivant les tranches d'âge pour 4 sous-groupes différents :
   - des femmes belges
   - des hommes belges
   - des femmes étrangères
   - des hommes étrangers

   Voici un échantillon de 3 communes du fichier *popage98.xls* :

	POP_1	POP_2	POP_3	$\cdots$	POP_{20}	SEXE	BELETR
Beauvechain	13	215	211	$\cdots$	2	1	1
Beauvechain	180	212	188	$\cdots$	4	2	1
Beauvechain	4	6	7	$\cdots$	0	1	2
Beauvechian	2	9	5	$\cdots$	0	2	2
Braine-L'alleud	1009	1056	1097	$\cdots$	3	1	1
Braine-L'alleud	881	1097	1058	$\cdots$	25	2	1
Braine-L'alleud	41	61	69	$\cdots$	2	1	2
Braine-L'alleud	47	50	63	$\cdots$	0	2	2
Mons	2320	2535	2348	$\cdots$	12	1	1
Mons	2237	2461	2198	$\cdots$	59	2	1
Mons	326	369	401	$\cdots$	1	1	2
Mons	312	382	378	$\cdots$	3	2	2

5. Le fichier *revenu.xls* procure une ventilation des déclarations fiscales par tranche. La variable décrivant la répartition des déclarations est constituée de 6 catégories :

   - V_1 : Le nombre de déclarations de moins de 100.000 FB.
   - V_2 : Le nombre de déclarations entre 100.000 FB et 250.000 FB.
   - V_3 : Le nombre de déclarations entre 250.000 FB et 500.000 FB.
   - V_4 : Le nombre de déclarations entre 500.000 FB et 700.000 FB.
   - V_5 : Le nombre de déclarations entre 700.000 FB et 1.000.000 FB.
   - V_6 : Le nombre de déclarations de plus de 1.000.000 FB.

   Voici un extrait du fichier *revenu.xls* :

	V_1	V_2	V_3	V_4	V_5	V_6
Beauvechain	84	141	426	468	490	974
Braine-L'alleud	547	811	2620	2539	2581	5970
Mons	1023	2067	8129	7830	7704	10047

Modification des fichiers

Pour utiliser correctement le programme, nous avons modifié certains fichiers :

1. Le fichier *agelo.xls*
   Nous avons supprimé dans le fichier *agelo.xls* la catégorie C_9. Cette catégorie représente le nombre de logements pour lesquels l'année de construction n'a pas été spécifiée.
   Etant donné que l'algorithme ne requiert que des modalités ordonnées, il est impossible de faire intervenir la catégorie C_9 dans l'analyse.

2. Le fichier *tailmen.xls*
   Les catégories T_1 et T_2 du fichier *tailmen.xls* donnent respectivement le nombre de ménages composés d'hommes seuls et de femmes seules.
   Comme il n'exite pas d'ordre sur ces deux catégories, nous avons créé une nouvelle catégorie qui représente le nombre de ménages composés d'une seule personne. Ainsi, l'ensemble des huit catégories sont bien ordonnées de façon croissantes. Pour chaque commune, le poids alloué à la nouvelle variable est la somme des poids de T_1 et de T_2.

   Exemple :

	T_1	T_2	La nouvelle catégorie
Beauvechain	167	228	$167 + 228 = 395$
Mons	5712	8098	$5712 + 8098 = 13810$

   Pour plus de facilités, nous avons appelé cette nouvelle catégorie T_2.

   Pour éviter de calculer, à la main, les poids de la nouvelle catégorie, nous avons écrit un programme en Fortran 77 appelé *som1.for* :

```
      PROGRAM SOMME1

c-------------------------------------------------------
c   Nous sommons les deux premieres colonnes T1 et T2 du
c    fichier tailmen.xls
c-------------------------------------------------------

c-------------------------------------------------------
c                declarations des variables
c-------------------------------------------------------
```

```
      IMPLICIT NONE
      INTEGER INS,J,S,I
      INTEGER V(10)
      CHARACTER*5 MOT(10)


c-------------------------------------------------------------
c                  corps du programme
c-------------------------------------------------------------

c ouverture des fichiers

      OPEN(11,FILE='tailmen.txt',STATUS='old')
      OPEN(12,FILE='essai2.txt',STATUS='new')

c lecture de la premire ligne, noms des variables

      READ(11,*) (MOT(I), I=1,10)
      WRITE(12,*) MOT(1),MOT(3),MOT(4),MOT(5),MOT(6),MOT(7),MOT(8),
     1     MOT(9),MOT(10)

c lecture des donnees et somme de deux colonnes

      DO 10 J=1,262

         READ(11,*) (V(I), I=1,10)
         S=V(2)+V(3)
         WRITE(12,*) V(1),S,V(4),V(5),V(6),V(7),V(8),V(9),V(10)

 10   CONTINUE

      CLOSE(12)
      CLOSE(11)

      END
```

3. Le fichier *popage98.xls*

Dans le fichier *popage98.xls*, la population de chaque commune est divisée en 4 sous-groupes :

- les hommes belges.
- les femmes belges.
- les hommes étrangers.
- les femmes étrangères.

Chaque sous-groupe est décrit par :

- les catégories POP_1, POP_2,..., POP_{20}.
- la variable $SEXE$ qui représente le sexe des personnes.
- la variable $BELETR$ qui désigne la nationalité des personnes.

Comme notre analyse porte sur l'entièreté de la population de chaque commune, nous avons additionné les quatre sous-groupes et supprimé les variables $SEXE$ et $BELETR$.

Prenons un exemple :

	POP_1	POP_2	$\cdots$	POP_20	SEXE	BELETR
Beauvechain	13	215	$\cdots$	2	1	1
Beauvechain	180	212	$\cdots$	4	2	1
Beauvechain	4	6	$\cdots$	0	1	2
Beauvechian	2	9	$\cdots$	0	2	2

En additionnant les quatre sous-groupes, nous obtenons :

	POP_1	POP_2	$\cdots$	POP_20
Beauvechain	199	442	$\cdots$	6

Pour réaliser cette addition, nous avons écrit un programme en fortran 77 appelé *som2.for* :

```
      PROGRAM SOMME2

c-------------------------------------------------
c   Nous sommons les quatre sous-groupes :
c           - les hommes belges.
c           - les femmes belges.
c           - les hommes etrangers.
c           - les femmes etrangeres.
c    pour obtenir un groupe global
c-------------------------------------------------
c-------------------------------------------------


c-------------------------------------------------
c           declarations des variables
c-------------------------------------------------

      IMPLICIT NONE
      INTEGER I,J,H,K
      INTEGER POP(21),P(21)
      CHARACTER*5 MOT(21)


c-------------------------------------------------
c           corps du programme
c-------------------------------------------------

c     ouverture des fichiers

      OPEN(11,FILE='popage98.txt',STATUS='old')
      OPEN(12,FILE='essai.txt',STATUS='new')

c     lecture de la premiere ligne, noms des variables

      READ(11,*) (MOT(I), I=1,21)
      WRITE(12,*)(MOT(I), I=1,21)
```

```
c     lecture des donnees et somme des quatre lignes

      DO 10 J=1,262

         DO  30 I=1,21
            POP(I)=0
30       CONTINUE


         DO 20 H=1,4

            READ(11,*) (P(I), I=1,21)
            POP(1) = P(1)

            DO 40 K=2,21
               POP(K) = POP(K) + P(K)
40          CONTINUE

20       CONTINUE


         WRITE(12,*) (POP(I), I=1,21)

10    CONTINUE

      CLOSE(12)
      CLOSE(11)

      END
```

4. Le fichier *popact.xls*
   Le programme de classification symbolique n'accepte que des catégories ordonnées. Or, le fichier *popact.xls* ne contient que des catégories non ordonnées. Comme nous voulons, malgré tout, insérer ce fichier dans nos recherches, nous avons effectué une analyse factorielle sur l'ensemble des catégories du fichier.

A l'aide du logiciel SPSS, nous avons réalisé une analyse factorielle sur l'ensemble des données du fichier. Notre objectif est de projeter l'ensemble des catégories sur le premier axe factoriel et d'ordonner les différentes catégories suivant leurs projections c'est-à-dire les poids obtenus pour le premier axe.

Pour calculer les poids, nous avons utilisé la méthode en composantes principales expliquée au chapitre 7.

Examinons les poids obtenus pour les différentes catégories et pour le premier et le deuxième axe factoriel :

```
Factor Matrix:
```

	Factor 1	Factor 2
S1	.98911	-.09990
S11	.90560	.42174
S2	.98730	.00577
S3	.98396	-.07910
S5	.99221	-.10346
S6	.97869	-.02817
S7	.98750	-.05261
S8	.99460	.00335
S9	.97602	-.03291

Nous pouvons remarquer que la plupart des poids obtenus pour le premier axe sont très proches et donc, il nous est impossible d'ordonner adéquatement les catégories.
Pour remédier à ce problème, nous avons appliqué une rotation orthogonale sur les axes factoriels appelée *Varimax*.
Nous avons obtenu comme résultats :

```
VARIMAX   rotation   1 for extraction 1.
VARIMAX converged in 7 iterations.
```

Rotated Factor Matrix:

	Factor 1	Factor 2
S1	.72963	.53986
S11	.41271	.40029
S2	.53293	.69281
S3	.73847	.50578
S5	.69834	.58482
S6	.74742	.44770
S7	.58812	.67206
S8	.60300	.61969
S9	.51089	.73459

Les poids correspondants au premier axe factoriel sont beaucoup plus dispersés. Ainsi, nous pouvons classer les 9 catégories dans l'ordre croissant :

$$S_{11}, S_9, S_2, S_7, S_8, S_5, S_1, S_3, S_6$$

Remarque :
Lorsque nous avons appliqué la rotation oblique *Quartimax* aux axes factoriels, les poids correspondants au premier axe étaient très proches. Et donc, il était impossible de pouvoir réaliser un classement des catégories.

8.2 Les fichiers **.sds*

8.2.1 Introduction

Trois fichiers d'entrée sont nécessaire à l'application de l'algorithme de classification symbolique, notamment, un fichier **.sds*. Le fichier **.sds* est construit à partir des fichiers de données grâce au logiciel "Microsoft ACCESS" et au programme DB2SO.

8.2.2 Le programme DB2SO

Le programme DB2SO a été élaboré dans le cadre du projet SODAS pour résoudre le problème de formalisation des données stockées dans les bases de données externes dans la structure des objets symboliques.
Le logiciel "Microsoft ACCESS" permet de transformer un fichier **.xls* en une base de données adéquate et de construire les "assertions" nécessaires à l'utilisation du programme DB2SO. De plus, dans ce programme, l'utilisateur peut choisir le type d'objets symboliques qui sera produit.
La commande EXPORT de ce programme construit le fichier **.sds* correspodant au fichier **.xls*.
Il est également possible de rassembler plusieurs fichiers **.xls* et d'obtenir un seul fichier **.sds* correspondant à l'ensemble de ces fichiers **.xls*.
En effet, grâce à la commande JOIN, nous pouvons fusionner les bases de données et les "assertions" correspondantes obtenues grâce au logiciel "Microsoft ACCESS" pour chaque fichier **.xls* et ensuite, obtenir un fichier **.sds* global.

8.2.3 La description des fichiers **.sds*

Un fichier **.sds* est divisé en plusieurs blocs :

- Le bloc **CONTAIN** contient la liste de tous les blocs présents dans le fichier **.sds*.

- Le bloc **FILE** donne des informations sur l'origine du fichier **.sds*.

- Le bloc **HEADER** répertorie les variables suivant leur type et donne le nombre de variables de chaque type ainsi que le nombre d'individus dans le fichier de données.

- Le bloc **INDIVIDUAL** assigne à chaque individu un numéro et un libellé.
  Exemple :
  Le numéro et le libellé de la commune de Marche-en-Famenne sont respectivement 22 et AA22.

- Le bloc **VARIABLE** assigne à chaque variable un numéro et un libellé.

- Le bloc **RECTANGLE-MATRIX** contient les caractéristiques de chaque individu. Chaque ligne de cette matrice représente les valeurs prises par les variables pour un individu.

8.3 Les fichiers d'entrée et de sortie

Comme nous l'avons rappelé à plusieurs reprises, la méthode de classification symbolique nécessite trois fichiers d'entrée :

1. Un fichier **.sds* créé grâce au logiciel Microsoft ACCESS et au programme DB2SO à partir des fichiers de données **.xls*.
2. Un fichier **.selec* créé par l'utilisateur. La première ligne de ce fichier donne le nombre de variables qui vont être traitées. La deuxième ligne cite les numéros des variables qui seront traitées.
3. Un fichier **.resu* vide créé par l'utilisateur.

Lors de l'utilistion du programme, l'utilisateur doit également choisir parmi trois distances pour calculer la matrice de dissimilarités :

1. La distance de Hausdorff non normalisée.
2. La distance de Hausdorff normalisée par l'inverse de la variance.
3. La distance de Hausdorff normalisée par l'inverse de l'écart maximum.

La méthode de classification construit un ensemble de partitions organisées dans un arbre binaire.
Le fichier **.resu* contient :

- L'arbre de division qui explique comment sont divisés les groupes et par quelles variables ou valeur de coupure.
- Les partitions progressives ainsi que les listes des individus présents dans chaque groupe des différentes partitions.

Chapitre 9

Les résultats

9.1 Recherches réalisées à partir du fichier *communes.xls*

Le fichier *communes.xls* comprend 262 communes décrites par 12 variables. Pour appliquer l'algorithme de classification, nous avons besoin de trois fichiers d'entrée :

1. *communes.sds* (cf. annexe A)

   Ce fichier a été créé à partir du fichier *communes.xls* grâce au programme DB2SO.

2. *communes.selec*

   Ce fichier donne le nombre et les numéros des variables traitées. Comme nous l'avons déjà dit , seulement 8 variables sont traitées : $VARPOP$, $POP75$, $MINIMEX$, $TAIL.MEN$, $BATI.LOG$, $CONFORT$, $OFFRE.EMPLOI$, $POP.ACT$.

   Communes.selec :

```
8
2  5  6  7  8  9  10  11
```

3. *communes.resu*

   Ce fichier est vide.

Comme nous sommes dans le cas de variables quantitatives, nous avons le choix entre trois distances pour calculer la matrice de dissimilarités :

- La distance de Hausdorff non normalisée.
- La distance de Hausdorff normalisée par l'inverse de la variance.
- La distance de Hausdorff normalisée par l'inverse de l'écart type.

Nous avons appliqué le programme à chaque fois avec une distance différente. Pour ces trois résultats, nous avons choisi cinq comme nombre de classes. Les trois fichiers résultat ont été soumis au Service de Démographie de la Région Wallonne.
Les résultats obtenus pour les trois distances n'ont rien apporté de consistant. Etant donné l'incohérence des résultats, nous avons préféré ne pas les joindre à ce mémoire.

9.2 Recherches réalisées à partir des fichiers *agelo.xls, popact.xls, revenu.xls, popage98.xls, tailmen.xls*

Grâce à la commande JOIN du programme DB2SO, nous avons rassemblé les cinq fichiers **.xls* et constitué un seul fichier *wal.sds* regroupant les cinq variables modales (cf annexe B).

Nous avons créé un fichier vide appelé *wal.resu* et un fichier *wal.selec* qui donne le nombre et les numéros des variables traitées :

wal.selec :

```
5
1  2  3  4  5
```

Comme nous sommes dans le cas de variables qualitatives ordonnées, nous ne disposons que d'une seule distance pour calculer la matrice de dissimilarités. Nous avons choisi de partitionner l'ensemble des 262 communes en au plus 7 classes. Les résultats ont été récoltés dans le fichier *wal.resu* (cf annexe C).

9.3 Interprétation des résultats du fichier *wal.resu*

9.3.1 Rappels : Division des classes

Soit C_i = ensemble de n_i objets.
Le but est de trouver la bipartition $C_i = (C_i^1, C_i^2)$ qui minimise la variance intra-classe. Nous choisissons la meilleure partition parmi toutes les partitions induites par l'ensemble de toutes les questions binaires possibles.

Questions binaires et données symboliques

- Un cluster C est divisé selon une question binaire de la forme :

$$Is\ \ Y_j \leq c\ \ ?$$

où $c \in \mathcal{Y}_j$ est appelé **valeur de coupure**

- Un objet $k \in C$ répond "oui" ou "non" à la question binaire selon une fonction binaire :

$$q_c : \Omega \rightarrow \{true, false\}$$

- Une bipartition (C_1, C_2) de C induit par une question binaire "$Is\ \ Y_j \leq c\ \ ?$" est comme suit :

$$\begin{aligned} C_1 &= \{k \in C \mid q_c(k) = true\} \\ C_2 &= \{k \in C \mid q_c(k) = false\} \end{aligned}$$

- Si Y_j est une variable modale telle que $Y_j(k) = \pi_k$, alors la fonction q_c est définie comme :

$$q_c(k) = \begin{cases} true & \text{si } \sum_{x \leq c} \pi_k(x) \geq \frac{1}{2} \\ false & \text{si } \sum_{x \leq c} \pi_k(x) < \frac{1}{2} \end{cases}$$

9.3.2 Première division

Considérons l'ensemble G des 262 communes de Wallonie.
Appliquons l'algorithme à cet ensemble de données.
Le groupe G est divisé en deux groupes G_1 et G_2 par la variable C_3 appartenant au fichier *agelo.xls*.

Rappels

Prenons deux exemples d'affectation :
Rappelons le critère :

$$q_c(k) = \begin{cases} true & \text{si } \sum_{x \leq c} \pi_k(x) \geq \frac{1}{2} \\ false & \text{si } \sum_{x \leq c} \pi_k(x) < \frac{1}{2} \end{cases}$$

	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Lens	759	100	101	34	121	40	23
Liege	16535	14795	17190	8476	7157	1034	581

1. Les catégories sont ordonnées de manière croissante. Une commune sera placée dans le premier groupe si la somme des poids des catégories inférieures ou égales à la variable de coupure est plus grande ou égale à la moitié de la somme des poids de toutes les modalités.

   Exemple : Lens est affecté dans le groupe G_1 :

   - Sommons tous les poids de toutes les catégories

     $$759 + 100 + 101 + 34 + 121 + 40 + 23 = 1178$$

   - Additionnons les poids des deux première catégories C_2 et C_3 :

     $$759 + 100 = 859$$

     Le résultat 859 est supérieur à la moitié de 1178 et donc Lens est bien placé dans le premier groupe.

2. Une commune sera placée dans le deuxième groupe si la somme des poids des catégories inférieures ou égales à la variable de coupure est plus petite que la moitié de la somme des poids de toutes les modalités.

Exemple : Liège est affecté dans le groupe G_2 :

- Sommons tous les poids de toutes les catégories :

$$16535 + 14795 + 17190 + 8476 + 7157 + 1034 + 581 = 65768$$

- Additionnons les poids des deux première catégories C_2 et C_3 :

$$16535 + 14795 = 31330$$

Le résultat 31330 est inférieur à la moitié de 65768 et donc Liège est bien placé dans le deuxième groupe.

Description

L'ensemble des 262 communes est divisé en deux groupes par la catégorie C_3 appartenant au fichier *agelo.xls*. La catégorie C_3 représente le nombre de logements construits entre 1919 et 1945. Les communes classées dans le groupe 1 ont construit de nombreux logements avant 1945 et peu après cette date. Par contre, les communes classées dans le groupe 2 ont construit de nombreux logements après 1945 (après la deuxième guerre mondiale).
Nous pouvons remarquer sur la carte 1 que les communes appartenant au groupe 2 longent la frontière allemande ou se situent à la périphérie de Bruxelles.
Les communes du groupe 1 se situent dans les régions de Mons, Tournai, Huy, Wavre et dans la région ardennaise.

9.3.3 Deuxième division

Description

Le critère a choisi de diviser le groupe G_1 en deux groupes (G'_1, G_3).
La catégorie de séparation est C_2 et appartient au fichier *agelo.xls*. Cette

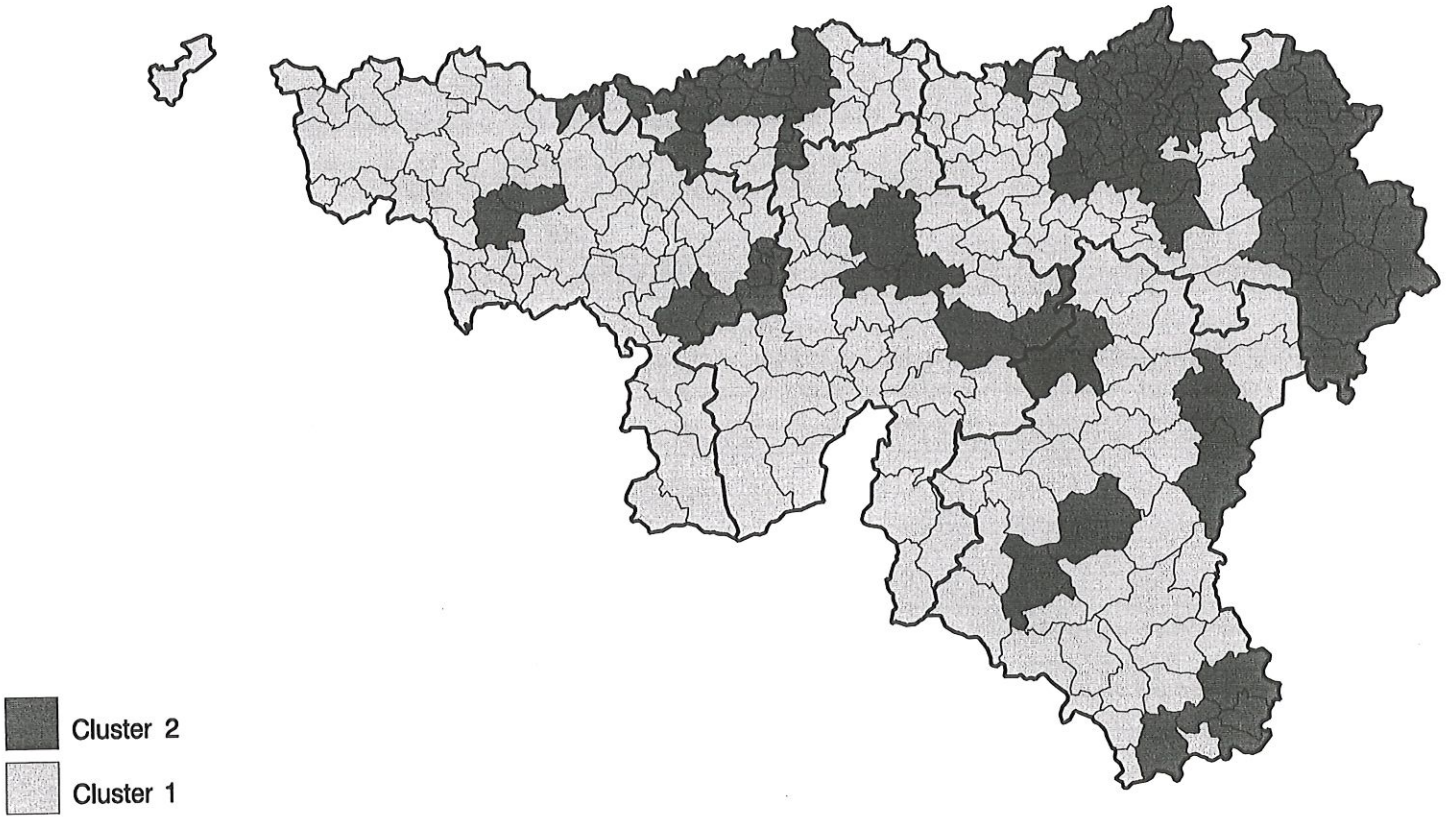
catégorie représente le nombre de logements construits avant 1919.
Les communes classées dans le groupe G_1' ont construit de nombreux logements avant 1919 et les communes du groupe G_3 ont construit beaucoup de logements entre les deux guerres 14-18 et 40-45.

9.3.4 Interprétations et critiques

Comparaison de la carte 3 avec la carte A :
Si nous comparons les communes du groupe G_2 avec les communes de la carte A dont l'âge moyen des logements est d'au plus 50 ans,
le groupe G_3 avec les communes dont l'âge moyen des logements se trouve entre 50 et 65 ans,
et le groupe G_1' avec les communes de plus de 65 ans, nous pouvons remarquer que nous retrouvons, à quelques exceptions près les mêmes communes sur les deux cartes. Notre comparaison est approximative et doit être prise avec beaucoup de précaution puisque la carte A date de 1991.

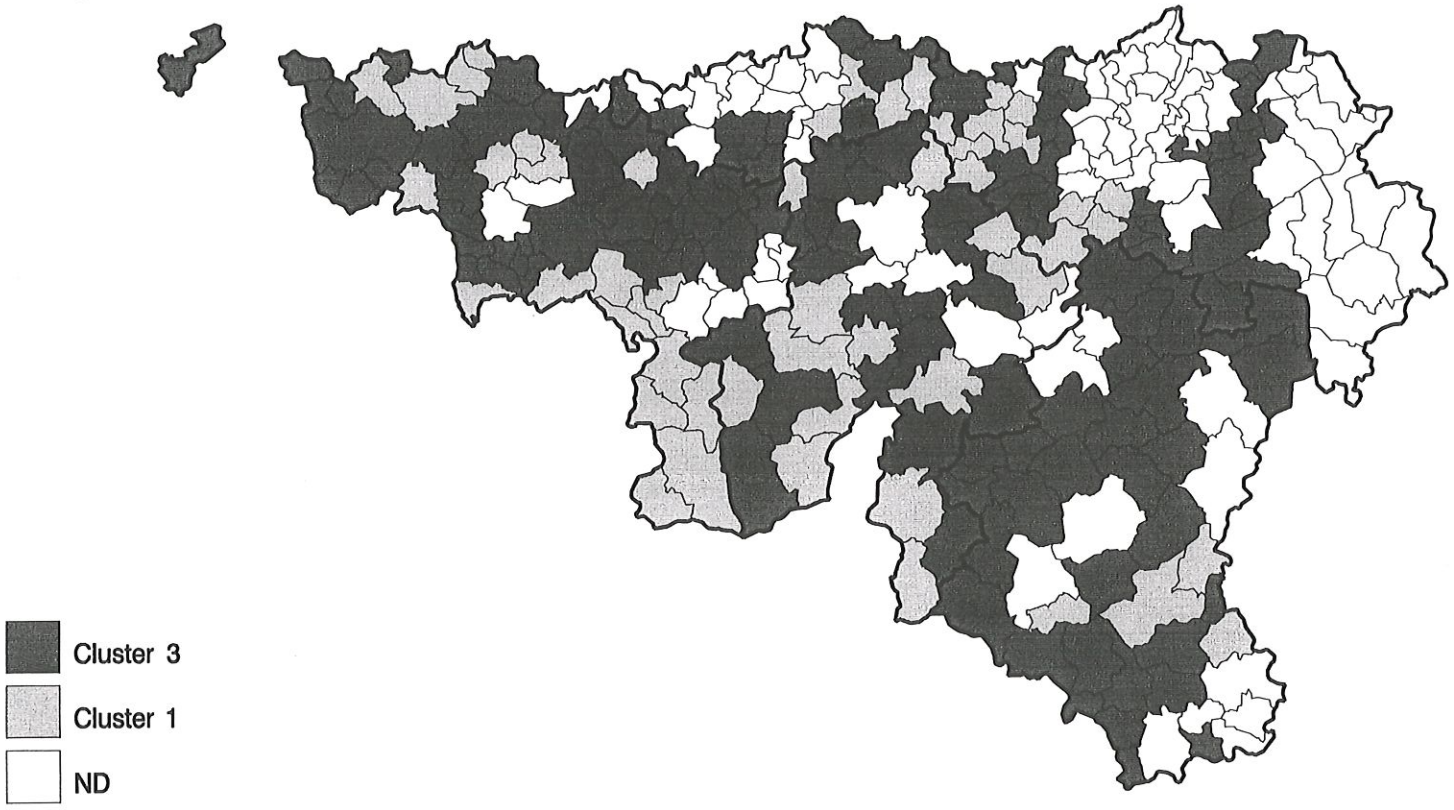
Interprétation :
La surreprésentation relative des anciennes habitations semble caractériser principalement les communes wallonnes. Il va de soi que la proportion des habitations très anciennes s'est réduite dans les régions ayant connu une forte activité de construction au XX siècle, et certainement après la seconde guerre mondiale dans les régions marquées par une forte démolition, comme c'est le cas dans la zone du front de la première guerre mondiale. Il est clair qu'une grande partie de la Wallonie comme la Fagne-Fammenne et l'Ardenne à savoir la zone située au sud de ce qu'on entend traditionnellement par l'axe industriel wallon, n'est pas aussi submergée par les habitations très anciennes. Il est également manifeste que les zones ayant une industrialisation précoce portent toujours l'héritage de leur passé, à savoir un grand nombre d'habitations anciennes. Le parc des logements est, dans une large mesure, ancien : un tiers des habitations ont plus de 50 ans et une bonne moitié des habitations ont été construites avant 1945.

Carte 1 : Division du groupe G en deux groupes G_1 et G_2



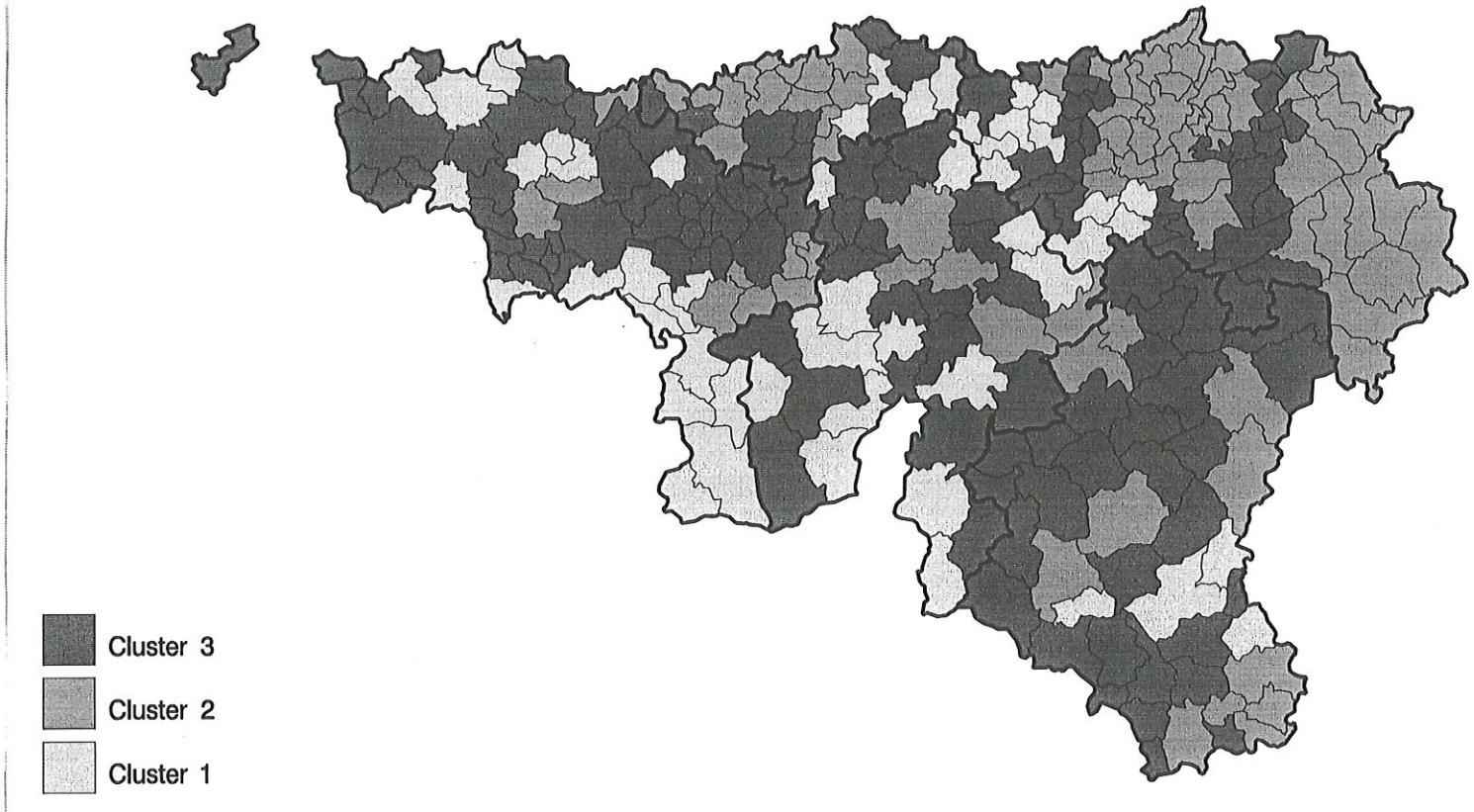
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 2 : Division du groupe G_1 en deux groupes G_1' et G_3



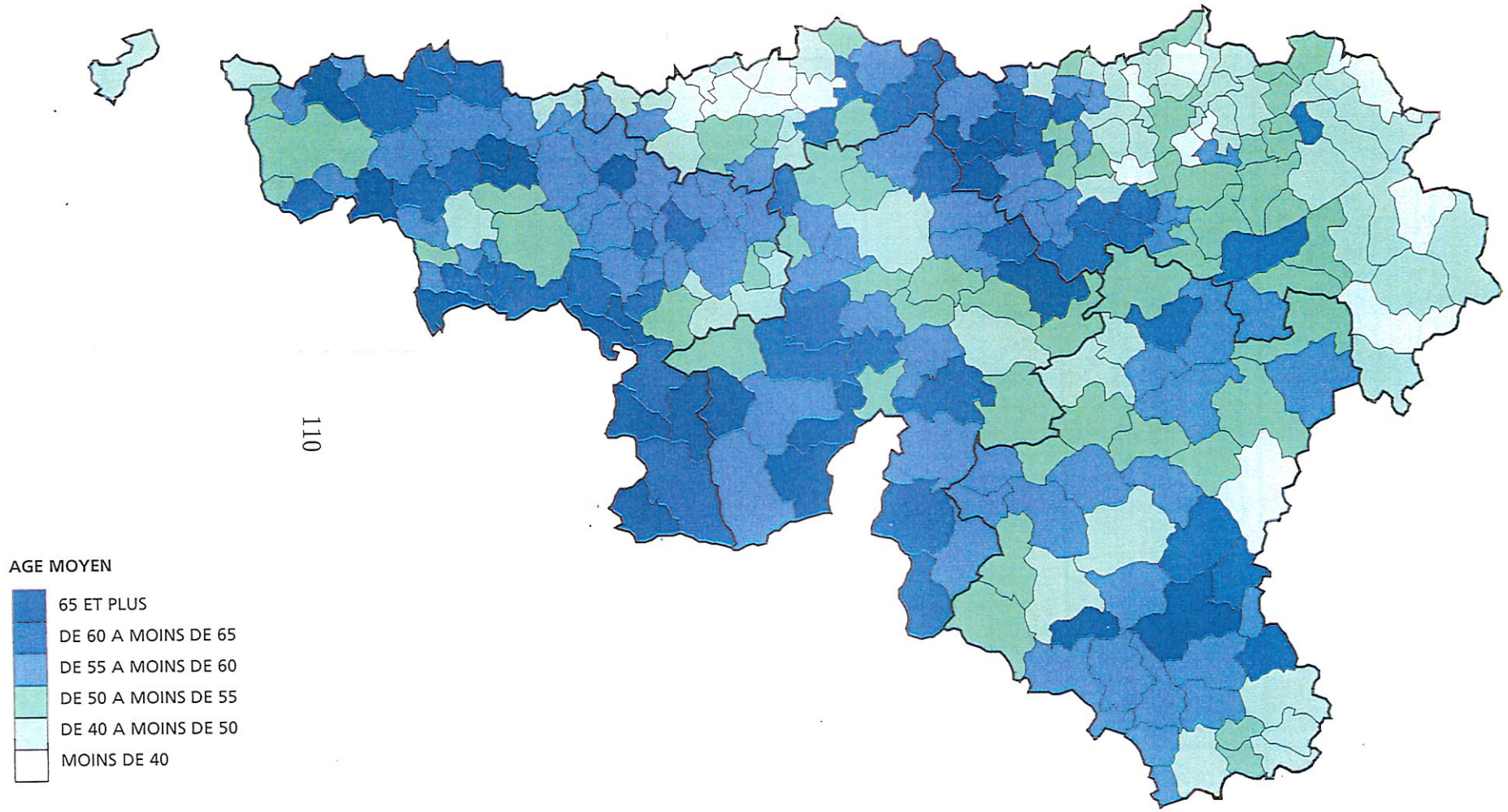
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 3 : Répartition en 3 clusters



SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte A : Age moyen des logements au recensement de 1991



SOURCE : Institut National de Statistique : "Recensement de la population 1991"
CALCULS : SES

9.3.5 Troisième division

Description

Le groupe 3 est divisé en deux groupes (G'_3, G_4) par la catégorie V_4 qui appartient au fichier *revenu.xls*. Cette catégorie V_4 représente le nombre de déclarations de revenus entre 500.000 FB et 700.000 FB.
Les communes qui ont construit de nombreux logements entre les deux guerres mondiales sont réparties dans deux groupes :

- Dans le groupe G'_3, la plupart de habitants de chaque commune ont des revenus moyens c'est-à-dire en dessous de 700.000 FB par an.
- Dans le groupe G_4, les revenus de la plupart des habitants de chaque commune sont assez élevés c'est-à-dire au-dessus de 700.000 FB par an.

Lorsque nous regardons le carte 4, nous pouvons remarquer que les villes où la population est la plus "pauvre" sont Tournai, Mons et Charleroi. Nous savons que la province du Hainaut a de sérieux problèmes économiques et les résultats le prouvent. Les communes qui regroupent les populations les plus riches sont situées à la périphérie de Bruxelles.

9.3.6 Quatrième division

Description

Le groupe 2 est également divisé en deux groupes (G'_2, G_5) par rapport à la variable V_4.
Rappelons que le groupe G_2 contient les communes qui ont construit de nombreux logements après la deuxième guerre mondiale. Comme dans le cas précédent :

- Dans le groupe G'_2, la plupart des habitants de chaque commune ont des revenus moyens c'est-à-dire en-dessous de 700.000 FB.
- Dans le groupe G_5 , la plupart des habitants de chaque commune ont des revenus élevés c'est-à-dire au-dessus de 700.000 FB.

Comme nous pouvons le voir sur le carte 6, les communes de ce groupe se situent dans la périphérie de la capitale ou se trouvent dans la région d'Arlon. Nous pensons que la plupart des travailleurs diplômés qui travaillent

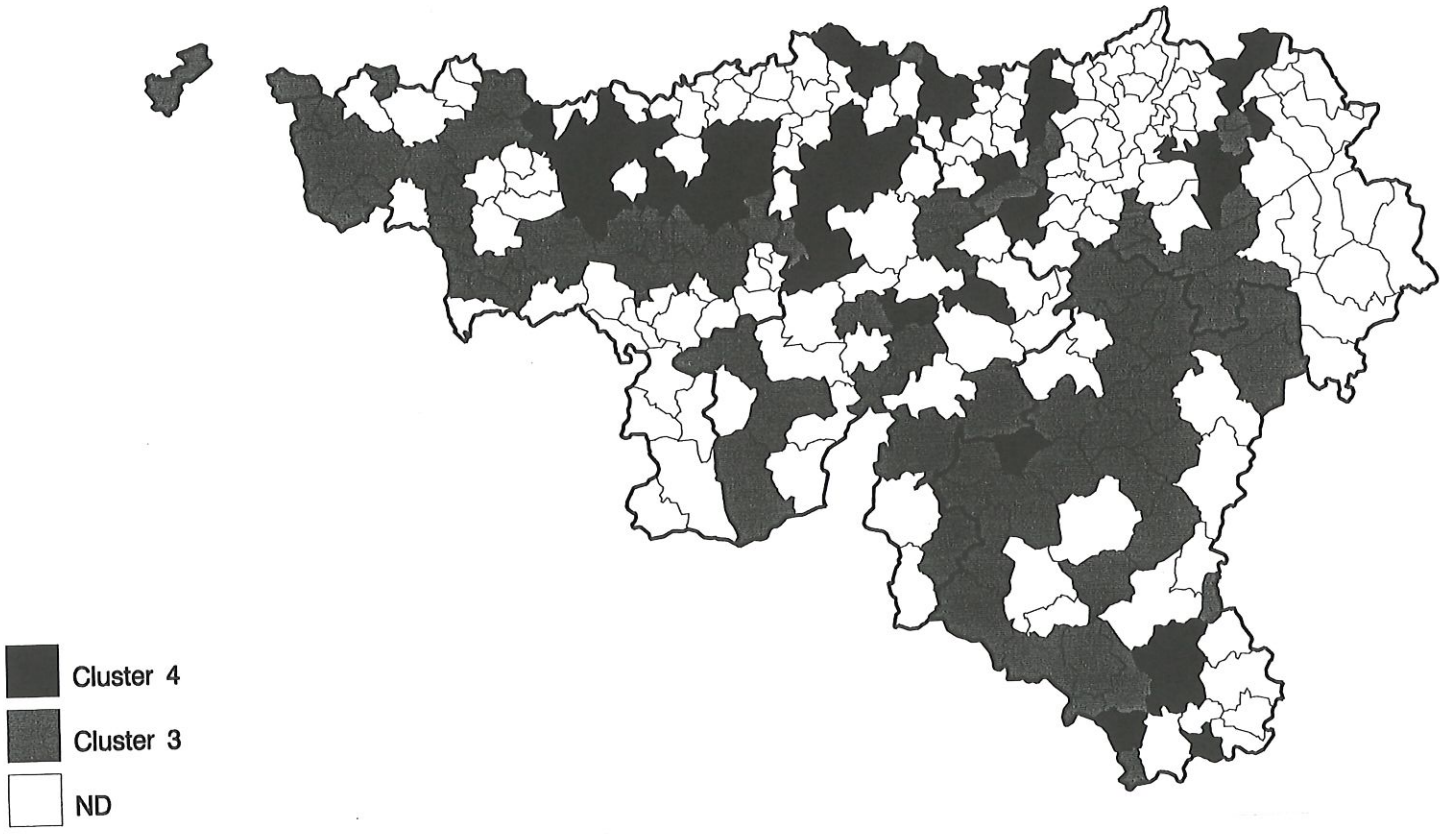
à Bruxelles habitent dans la périphérie de la capitale tandis que les Belges qui travaillent au Luxembourg et qui touchent de très bons salaires habitent dans la région d'Arlon.

9.3.7 Interprétations et critiques

Comparaison de la carte 7 avec la carte B :
Si nous comparons les communes des groupes G_2' et G_3' dessinées sur la carte 7 avec les communes de revenu médian d'au plus 700.000 FB de la carte B et les communes des groupes G_4 et G_5 dessinées sur la carte 7 avec les communes de revenus médians d'au moins 700.000 FB de la carte B alors nous pouvons remarquer que nous retrouvons plus ou moins les mêmes communes dans les différents groupes. Mais, il faut prendre ces conclusions avec beaucoup de précautions car la carte B est une carte des revenus médians de 1997.

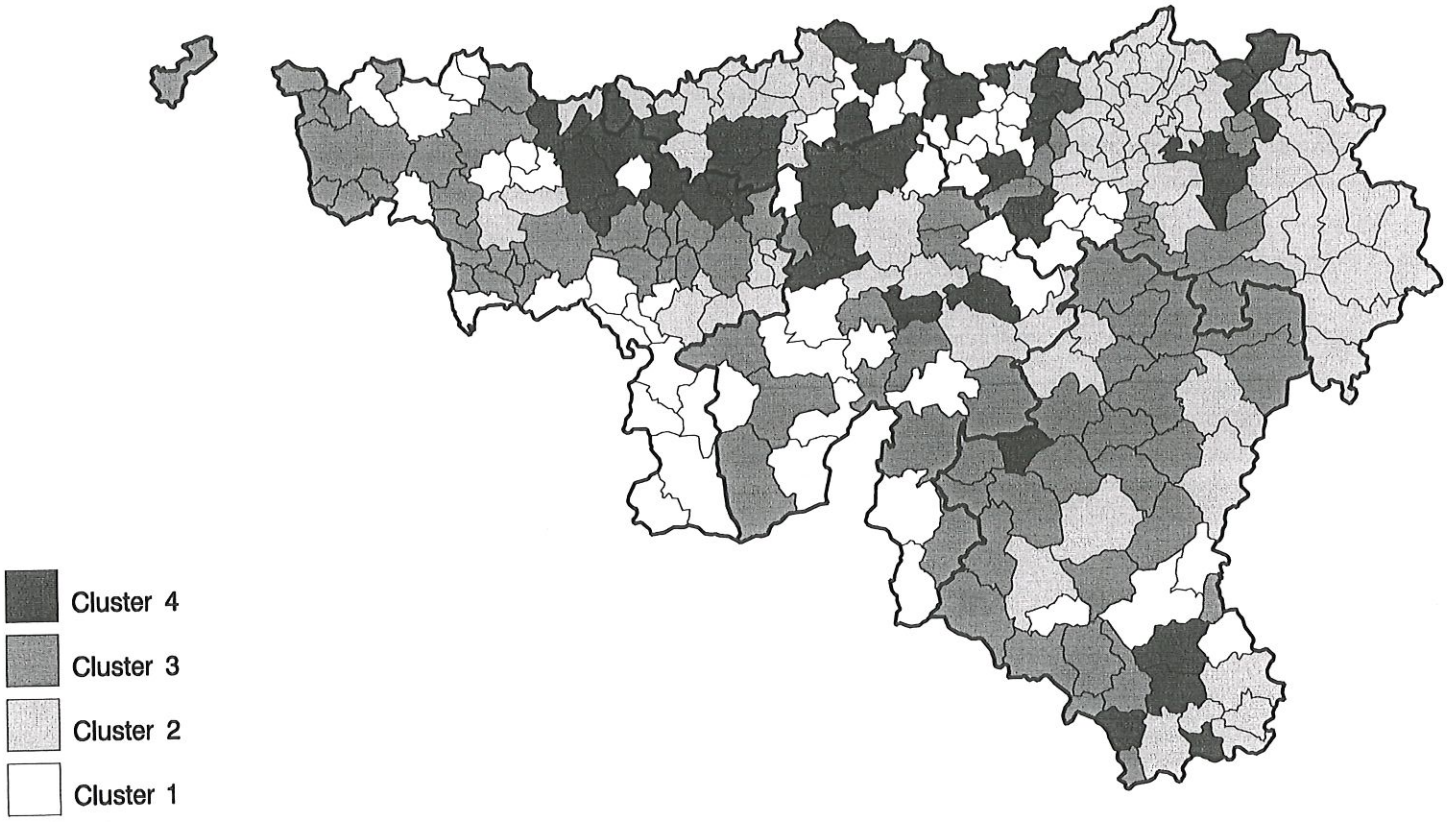
Interprétation :
On remarque en l'occurence que la ville "centre" dispose d'un revenu moindre que celui d'une série de communes agglomérative ou de banlieue de la même région urbaine. La classe des revenus supérieurs est située surtout à la périphérie de Bruxelles. Ainsi, la plupart des communes du centre du pays avec entre autres la région urbaine de Louvain et le Brabant wallon appartiennent aux deux catégories les plus élevées. Les revenus les plus faibles sont relevés dans plusieurs communes de l'ancien axe industriel wallon, plus spécifiquement en Hainaut, où le vieillissement et le chômage ont atteint des valeurs très élevées. Le caractère rural affirmé associé à la forte proportion d'agriculteurs parmi la population active explique la faiblesse des revenus relevés sur des zones importantes comme le sud-est de la Belgique. Des valeurs supérieures apparaissent dans le sud du Luxembourg. Ceci s'explique par le nombre important de travailleurs occupés dans la ville de Luxembourg. La pauvreté touche d'abord les jeunes ménages et les classes qui se trouvent en fin de vie active.

Carte 4 : Division du groupe G_3 en deux groupes G'_3 et G_4



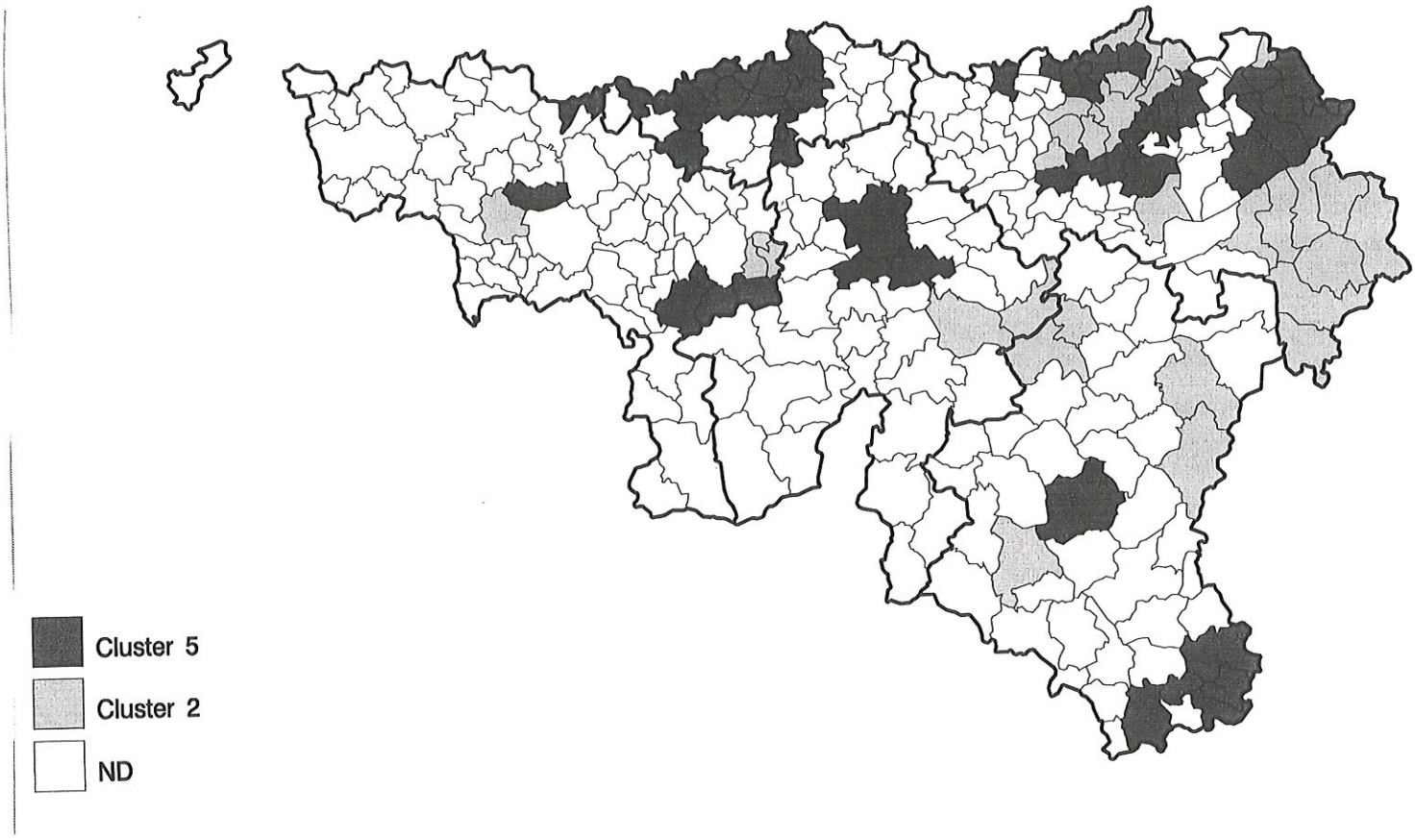
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 5 : Répartition en 4 clusters



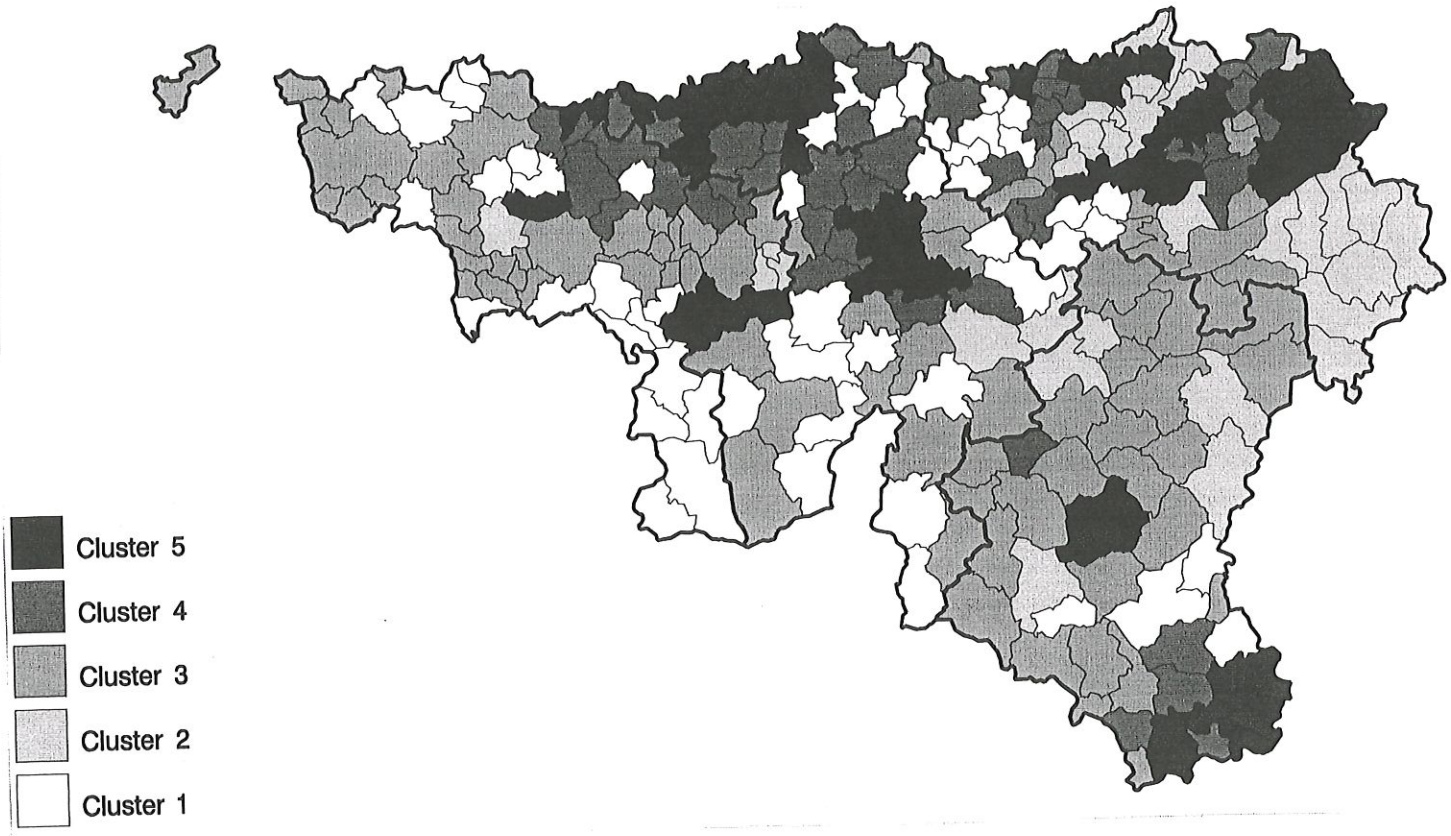
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 6 : Division du groupe G_2 en deux groupes G_2' et G_5



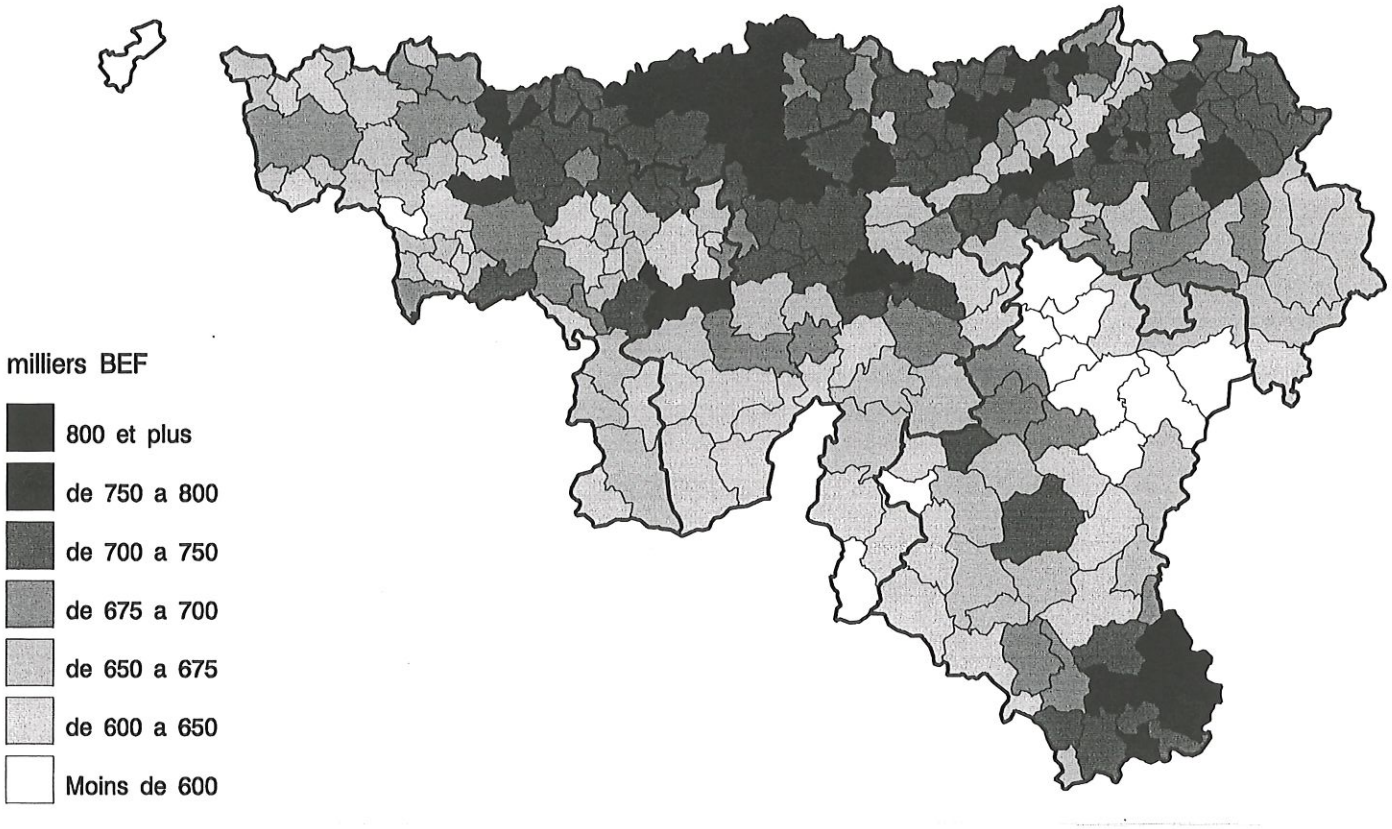
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 7 : Répartition en 5 clusters



SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte B : Revenu médian exercice fiscal 1997



SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

9.3.8 Cinquième division

Description

Le groupe G_2' est divisé en deux groupes (G_2'', G_6) par rapport à la catégorie T_3 qui appartient au fichier *Tailmen.xls* (voir carte 8).
Cette catégorie représente le nombre de ménages composées de 2 personnes.
Rappelons que le groupe G_2' contient les communes pour lesquelles les revenus des habitants sont peu élevés.

- Dans le groupe G_2'', la plupart des ménages des communes sont composés d'au plus deux personnes.
- Dans le groupe G_6, la plupart des ménages des communes sont composés d'au moins 3 personnes, malgré que les revenus de ces ménages soient peu élevés.

9.3.9 Sixième division

Description

Le groupe G_5 contient l'ensemble des communes qui vérifient la propriété suivante : "les revenus de la plupart des habitants sont élevés".
Le groupe 5 est divisé en deux groupes (G_5', G_7) par la catégorie T_3 également (voir carte 10).

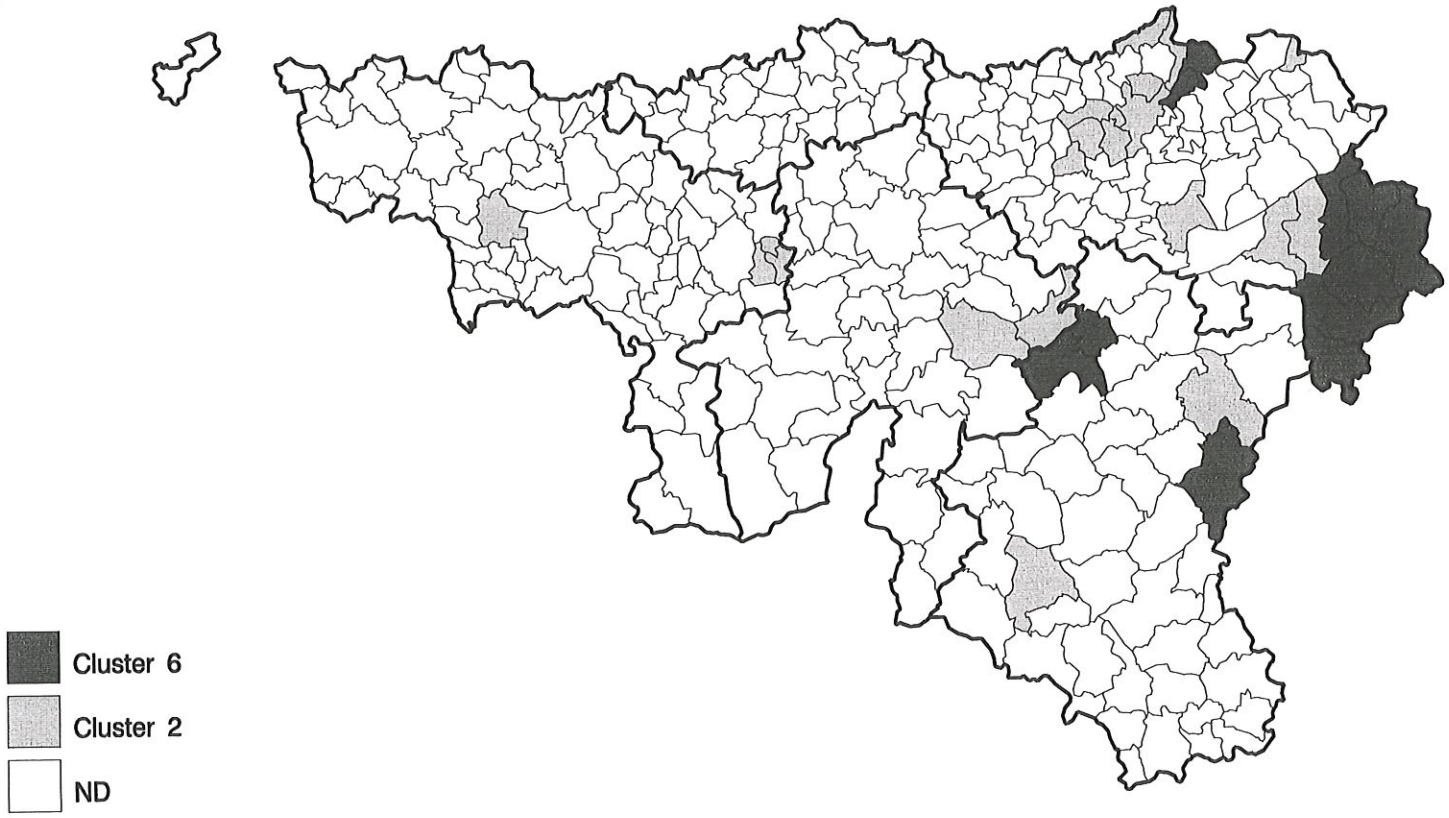
- Dans le groupe G_5', les tailles des ménages des communes sont pour la plupart inférieures ou égales à 2.
- Dans le groupe G_7, les tailles des ménages des communes sont pour la plupart supérieures à 2.

9.3.10 Interprétation et critiques

Comparaison :
La comparaison des cartes 8,10 et 11 est très difficile et non pertinente comme le nombre de commune considéré est trés faible. De plus, la plupart du temps, les démographes ne considèrent , dans l'analyse de composition des ménages que les cartes représentant les répartitions des personnes seules et les ménages composés d'au moins 5 personnes dans les différentes communes.

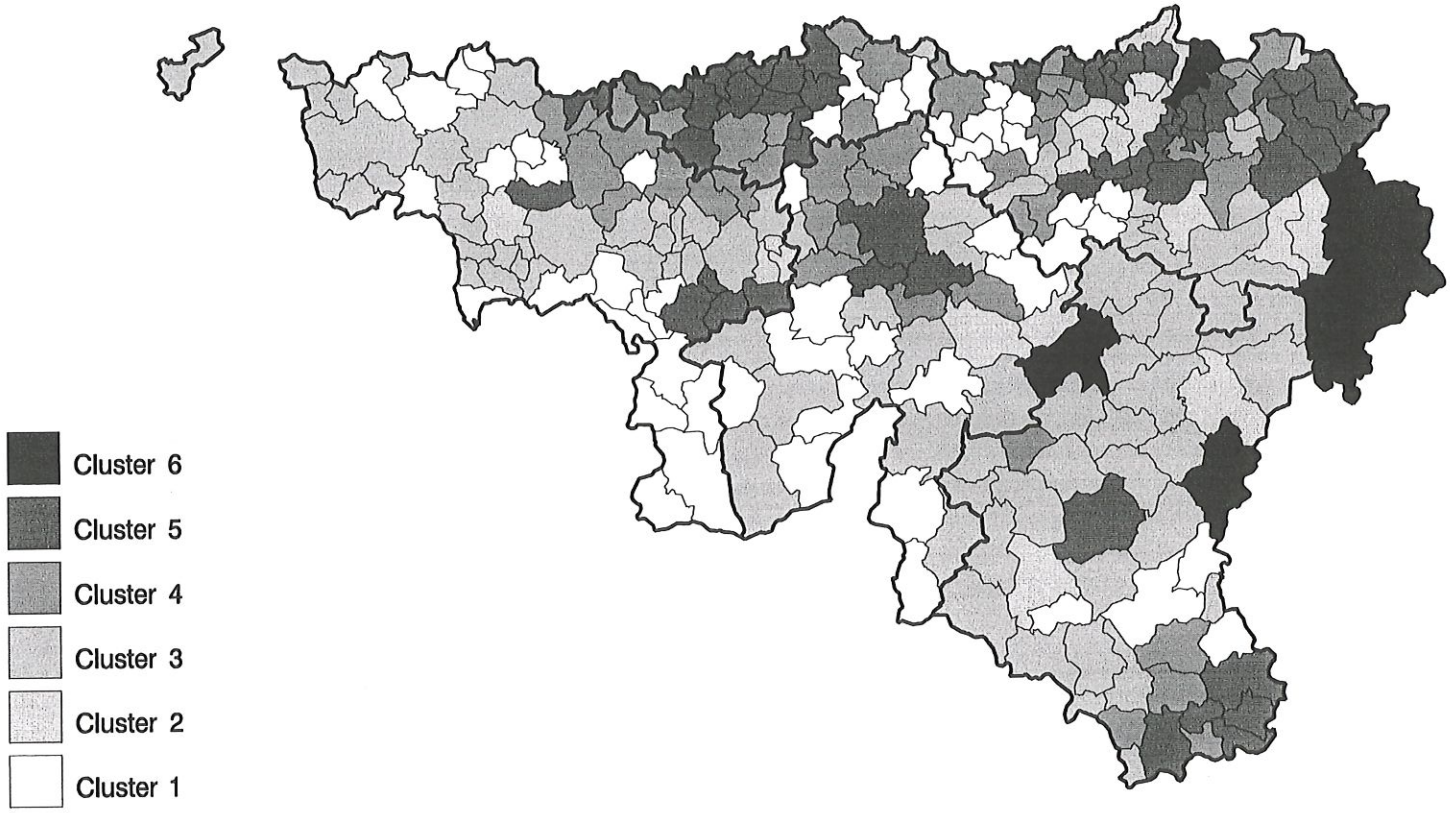
Interprétation :
La taille des ménages se calcule sur base du nombre de personnes séjournant (c'est-à-dire domiciliées légalement) dans ces ménages dit particuliers. Cela signifie donc que les personnes domiciliées dans les ménages collectifs (congrégations, prisons, maisons de repos) ne sont pas prises en considération. Les communes dont le nombre de ménages de taille égale ou inférieure à deux personnes sont Louvain, la région de Bruxelles-Capital, Liège et Arlon (groupes G_2'', G_5'). Les communes (groupes G_6, G_7) dont la taille des ménages est suprérieure à deux personnes sont les communes rurales ou les communes des banlieues résidentielles des villes. Les communes rurales de la province de Hainaut comptent en général des ménages d'une taille moyenne inférieure à celle des autres communes. De plus, il faut prendre en compte également que les jeunes couples attendent longtemps pour avoir des enfants.

Carte 8 : Division du groupe G_2' en deux groupes G_2'' et G_6



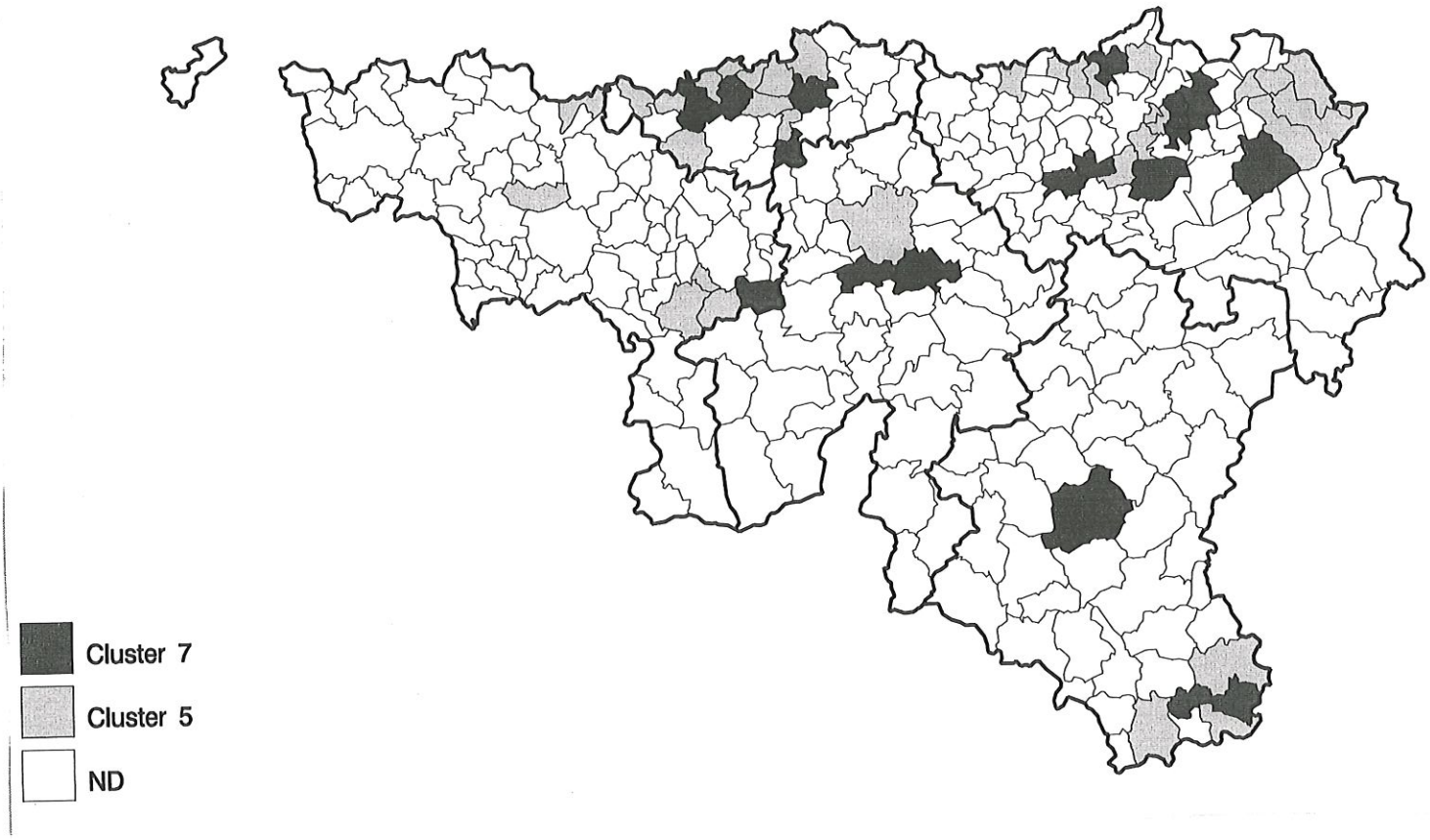
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 9 : Répartition en 6 clusters



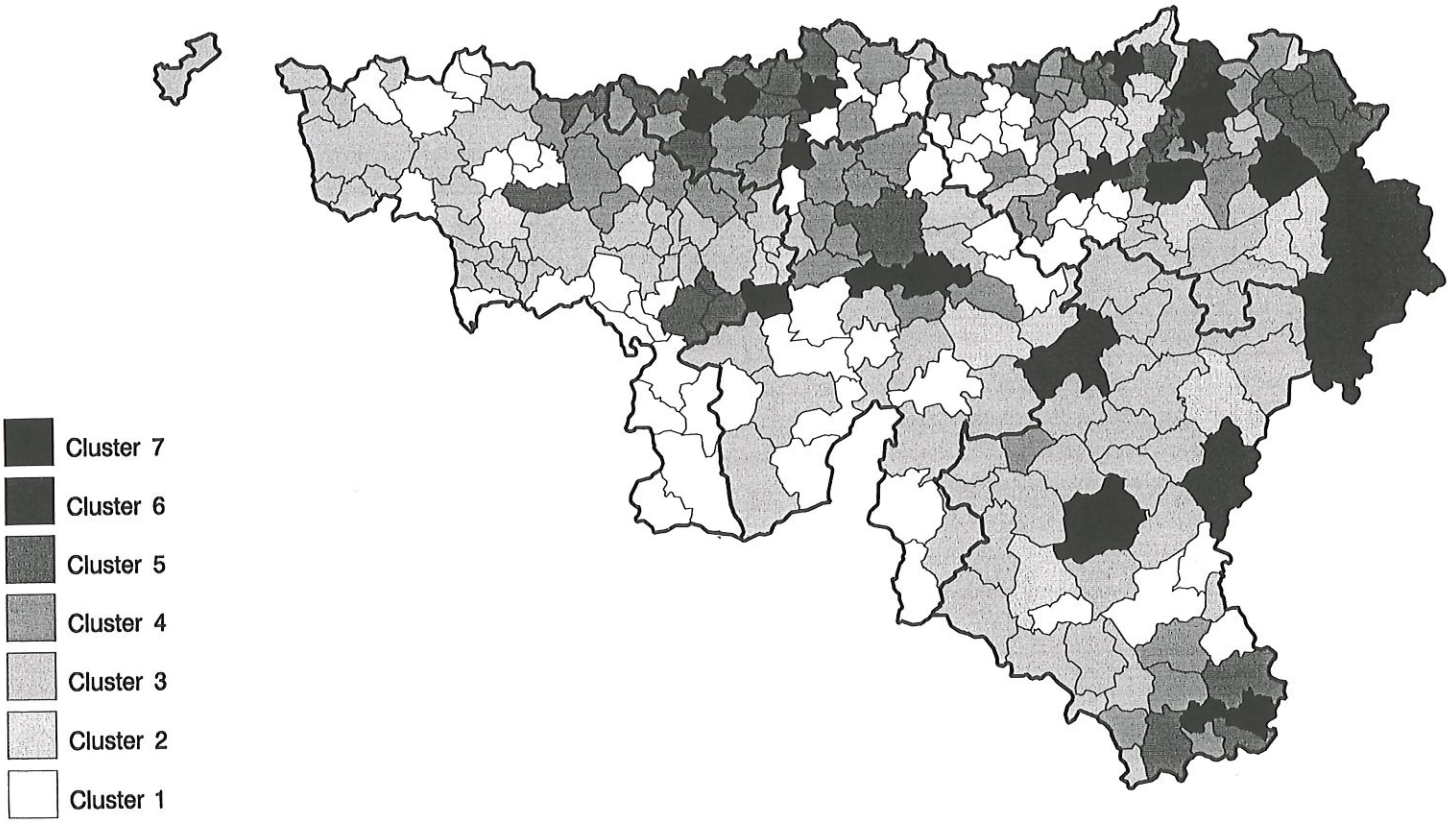
SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 10 : Division du groupe G_5 en deux groupes G_5' et G_7



SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Carte 11 : Répartition en 7 clusters



SOURCE ET DESIGN : Service des Etudes et de la Statistique (Région Wallonne)

Chapitre 10

Bugs

Lors de la réalisation de ce mémoire, nous avons rencontré de nombreuses difficultés :

1. Nous avons reçu les fichiers **.xls* assez tardivement. De plus, ces fichiers ont dû être modifiés pour pouvoir être utilisé dans l'algorithme de classification.

2. Lorsque nous avons utilisé le programme DB2SO pour transformer les bases de données en fichier **.sds*, les poids obtenus dans le bloc "RECTANGLE-MATRIX étaient incohérents.
   Nous avons dû renvoyer le programme DB2SO pour être corrigé.

3. Comme les résultats obtenus pour le fichier *communes.sds* n'étaient pas très pertinents, nous voulions essayer de modifier le critère. Mais, étant donné que le temps dont nous disposions était assez restreint et que de plus, le programme est implémenté en C_{++}, langage auquel nous n'avons jamais été initié, il nous était impossible de le modifier dans de bonnes circonstances.
   L'algorithme implémenté en C_{++} est divisé en 18 fichiers emboités qui sont disponibles sur internet *http ://www.-rocq.inria.fr/sodas/wp$_1$/sodas-paser* . De plus, ces fichiers ne contiennent aucun commentaire.

Conclusion

Ce mémoire a été consacré à l'application d'une méthode divisive monothétique de classification a des données modales et classiques. Nous avons appliqué cette méthode à deux jeux de données comprenant respectivement des types de variables différents.

Les résultats du premier fichier de données "communes.xls" n'étaient pas très pertinents. Alors, nous avons testé la méthode pour des données modales sur un ensemble de fichiers dont les variables décrivent les communes Wallonnes. L'application de cette méthode nous a permis de mettre le doigt sur certains problèmes et difficultés. Lorsqu'une commune est placée dans un groupe, elle y reste jusqu'à la fin. En d'autres termes, il est impossible de corriger une mauvaise partition. Le critère de séparation d'un cluster en deux clusters choisit arbitrairement le poids $\frac{1}{2}$ pour répartir les différentes communes dans les nouveaux clusters.

Nous avons réalisé grâce à l'aide du Service des Etudes et de la Statistique les cartes correspondant aux différentes coupures et aux différents clusters. Mais, le problème est qu'il n'existe pas toujours de cartes officielles correspondant aux cartes créées. Donc, notre interprétation reste vague et floue. De plus, les cartes officielles prennent en considération l'ensemble des 262 communes dans l'analyse. Comme la méthode de classification est une méthode hiérarchique, plus les clusters sont divisés, plus le nombre de communes prises en considération et représentées sur les cartes diminue.

Pour l'ensemble des quatres premières divisions, nous en concluerons que les résultats recoupent assez bien les cartes officielles mais pour les deux divisions suivantes, il est impossible d'interpréter convenablement les résultats comme le nombre de communes dans l'analyse n'est plus assez grand.

Il serait peut être intéressant de continuer ce travail en testant la méthode sur d'autres ensembles de données comme par exemple, les données de Ruspini. La méthode doit être approfondie et modifiée. Elle mérite une plus ample réflexion sur le critère de sélection et notamment sur la manière de choisir les "valeurs" de coupure.

Bibliographie

[1] H.H.BOCH *Scientific report Version E* , Elancourt, June 2, 1998.

[2] J.P. DUPREZ *Analyse multivariée de l'emploi dans les communes wallonnes*, Services des études et de la statistique, Namur, 1996.

[3] X.BARON, *Apport du son, de la couleur et de la 3D à la représentation des objets symboliques*, mémoire de l'institut d'informatique aux FUNDP, Namur, 1996.

[4] M. JAMBU, *Méthodes de base de l'analyse des données*, collection technique et scientifique de télécommunications, Eyrolles, Paris, 1999.

[5] M.CHAVENT, *Software requirements specification for the divisive clustering method*, prepared for SODAS, Lise-Ceremade university Paris IX-Dauphine, Paris, 1998.

[6] J.L.CHANDON et S.PINSON, *Analyse typologique : théorie et applications*, Masson, Paris, 1995.

[7] B.EVERITT, *Cluster Analysis*, Halsted press, London, 1980.

[8] X.BRY, *Analyses factorielles simples*, Economica, Paris, 1995.

[9] P.DAGNELIE, *Analyse statistiques à plusieurs variables*, Presses Agronomiques de Gembloux, Gembloux, 1975.

[10] A.C.RENCHER, *Methods of multivariate analysis*, Wiley, New-York, 1995.

[11] B.MERENNE, *La Belgique : Diversité territoriale*, Bulletin du Crédit Communal, Trimestriel, numéro 202, 1997.

[12] C. ALBASSART, *Annuaire statistique de la Wallonie 1997*, Services des études et de la Statistique, Namur, 1997.

[13] L'INSTITUT WALLON ASBL *Atlas de la Wallonie*, Ministère de la Région Wallonne, Namur, 1998.